Digital Humanities 2017

Conference Abstracts

McGill University & Université de Montréal

Montréal, Canada

August 8 – 11, 2017

Prepared by Rhian Lewis and the DH2017 Local Organizers:
Cecily Raynor, Dominic Forest, Michael Sinatra and Stéfan Sinclair

Program Committee

Diane Jackacki (Program Chair, CSDH)
Aurélien Berra (EADH)
Jeremy Boggs (ACH)
Marco Buechler (centerNet)
Johanna Drucker (ACH)
Dominic Forest (CSDH)
Asanobu Kitamoto (JADH)
Christian-Emil Smith Ore (EADH)
Laura Mandell (centerNet)
Sophie Marcotte (Humanistica)
Maki Miyake (JADH)
Simon Musgrave (AADH)
Glen Roe (AADH)

Local Organizers

Michael Sinatra
Stefan Sinclair
Cecily Raynor
Dominic Forest

Sponsors

Social Sciences and Humanities Research Council of Canada
Association of Digital Humanities Organizations
McGill University (Faculty of Arts)
Université de Montréal (Faculté des Arts et des Sciences)
Canadian Society for Digital Humanities / Societé canadienne des humanités numériques
Humanistica: association francophone des humanités numériques
Centre de Recherche Interuniversitaire sur les Humanités Numériques (Université de Montréal)
Kule Institute for Advanced Study (University of Alberta)

Welcome to Digital Humanities 2017

Table of Contents

Panels

Long Papers

Short Papers

Virtual Short Papers

Posters

Pre-Conference Workshops

SIG-Endorsed Pre-Conference Workshops

Panels and Multi-Paper Sessions

Accessing Alternative Histories and Futures: Afro-Latin American Models for the Digital Humanities

Eduard Arriaga
earriaga@alumni.uwo.ca
University of Indianapolis, United States of America

Andrés Villar
avillar2@uwo.ca
University of Western Ontario, Canada

Yvonne Captain
ycaptain@gwu.edu
George Washington University, United States of America

Maria Cecilia Martino
maricelmartino@gmail.com
Universidad de Buenos Aires, Argentina

Introduction

The field of the digital humanities has broadened substantially in the last few years, and this expansion has, in its turn, intensified debates about digital diversity and accessibility. These debates have emerged from the realization that, in spite of the euphoria created by visions of a "flat" digital landscape with access for all, the results to date do not measure up to such expectations. New associations with languages and academic methodologies that differ from those of the global North are promoting initiatives to investigate questions of access to technologies associated with the digital humanities. Despite these contributions by organizations such as the Post Colonial Digital Humanities (DHPoCo), FemTech.Net, Transform DH, GO:DH, Associaçao das Humanidades Digitais, etc., the issue of making the digital humanities more diverse and more accessible remains a problem of concern. Questions about race and ethnicity, and about the very material foundations —people, geographies, and tools— that make the digital possible are sometimes lost in discussions about how the digital domain is constituted and distributed across the globe. Therefore, the possibilities, or lack thereof, for taking advantage of digital resources and digital connectivity by people who have been, and continue to be, under-represented in the digital humanities remains to be a topic of some urgency. More conversations are needed across the digital humanities to address and bridge this gap, and to propose more productive ways of thinking about the impact, benefits, and drawbacks of the digital domain as it becomes an increasingly important layer of our cultural and social lives.

It is in the spirit of examining what we have been alluding to as a digital gap that our panel brings together specialists and practitioners of the digital humanities to discuss specific issues in countries and communities that have traditionally been at the periphery of normative cultural representations in various parts of the southern Americas. How are Afro-Latin American and Afrolatin@ communities using methodologies associated with the digital humanities to advance their own goals? How do current digital tools as applied to curatorial projects, archives, and the dissemination of information allow these same communities to explore hidden histories and propose new images and (self-) representations? What kind of digital resources —whether in terms of big-data or minimal computing— are these communities using to intervene in debates about the current state, and the future, of the digital domain? These are some of the fundamental questions inextricably tied to problems of accessibility that our panel will address by means of specific case studies. The diversity of methodologies we will bring to the conversation becomes a strength that productively challenges rigid limits on how the digital humanities are to be defined, which is particularly important when inclusion is at stake.

The session itself will consist of a brief introduction by the panel organizers explaining the rationale for the choice of papers and how these shed light on issues of accessibility in the digital humanities. This will be followed by three presentations, two of which will be in English and one in Spanish (María Cecilia Martino):

International Organization Theory (IO) and Online Afro-Latin America

Yvonne Captain

This presentation will discuss the connection between International Organization Theory, Digital Humanities and the connection of diverse Afro-Latin American political, social and cultural organizations. Professor Captain will show how, using digital tools to reach out to global communities, organizations such as Afroféminas [a Spanish/Latin American organization] help new identities emerge and connect productively. Her paper will hone in on the work of Afro-Costa Rican poet Shirley Barr Campbell, who connects to a variety of organizations via digital tools as a way to draw larger audiences and start discussing what it means to be a citizen of the Diaspora. Her paper focuses on how the advent of digital networks has impacted the formation, membership and digital appropriation of organization that have existed and connected well before our digital boom. In addition, she considers those organizations that have failed in connecting themselves digitally as examples that “proffer a cautionary lesson on what not to do” and as examples on how to handle digital waste.

Towards the Organization of an Afro-descendant Digital Archive in the Cape Verdean Association

María Cecilia Martino

This presentation focuses on the experience of an association of Cape Verdean migrants to Buenos Aires (Argentina) in organizing information by means of a digital archive. In her paper, Mrs. Martino discusses issues of access to technology, the creation of a basic system of metadata and the connection between members of the community based on materials such as photographs, videos, and documents that tell the story of migration and resilience of Cape Verdeans immigrants to Buenos Aires. Finally, her paper will discuss how digital methodologies would help enrich ethnographic practices concerned with representing people's discourse.

Afrolatin@ Digital Humanities

Eduard Arriaga

The final presentation discusses projects carried out by and about Afrolatin@ and Afro-Latin American communities in the Americas, questioning the very definition of Digital Humanities in order to include such initiatives. By analyzing the studied projects, this paper questions conceptions of humanity, humanities, access and digital appropriation residing at the core of current digital humanities projects and put to test by Afrolatin@ and Afro-Latin American endeavors. Finally, the paper reviews some of the principles of the minimal computing approach to identify its limits, advantages and impacts in Afrolatin@ digital projects.

Critical Digital Humanities and Machine Learning

Caroline Bassett
c.bassett@sussex.ac.uk
University of Sussex, United Kingdom

David M. Berry
d.m.berry@sussex.ac.uk
University of Sussex, United Kingdom

M. Beatrice Fazi
b.fazi@sussex.ac.uk
University of Sussex, United Kingdom

Jack Pay
jp242@sussex.ac.uk
University of Sussex, United Kingdom

Ben Roberts
b.l.roberts@sussex.ac.uk
University of Sussex, United Kingdom

Panel Statement

This panel undertakes a speculative and theoretical discussion of possible future directions for digital humanities work driven by what we call informating, augmenting and automating technologies in the digital humanities. The panel particularly examines the emergence of a new paradigm of artificial intelligence around machine learning, statistical techniques and textual interfaces; a paradigm that challenges the way in which we understand the provision of digital humanities technologies and infrastructures. We explore the debates over informating, augmenting and automating processes that are now starting to emerge in digital humanities, and the historical trajectory that led to the current rapid changes from computational techniques. By looking at how machine-learning infrastructures effect knowledge formations, we engage with these new knowledges and practices, and argue that digital humanities must seek to contest and transform particular institutional structures that are problematic for humanities scholarship.

Although differences have emerged within the digital humanities between "those who use new digital tools to aid relatively traditional scholarly projects and those who believe that DH is most powerful as a disruptive political force that has the potential to reshape fundamental aspects of academic practice" (Gold, 2012: x), it is still the case that, as a growing and developing disciplinary area, digital humanities has much opportunity for these disparate elements to work together. Not unlike differences between empirical and critical sociology in a previous iteration of a contestation over knowledge, epistemology, disciplinary identity and research, digital humanities as a discipline will be richer and more vibrant with alternative voices contributing to projects, publications and practices. Indeed, the debates within digital humanities "bear the mark of a field in the midst of growing pains as its adherents expand from a small circle of like-minded scholars to a more heterogeneous set of practitioners who sometimes ask more disruptive questions" (Gold, 2012: x-xi).

Developing a critical approach to machine learning, for example, calls for computation itself to be historicised, and its developing relationship with humanities to be carefully uncovered. Similarly, by focusing on the materiality of machine learning, our attention is drawn to the microanalysis required at the level of computational conditions of possibility, combined with a macroanalysis of deployment of machine-learning systems in humanities work. This calls for us to think critically about how machine learning is being designed and deployed in the specific problem domains represented by the informating, augmenting and automating of digital humanities. The panel critically engages with these three modes of thought and practice, in order to connect and explore the present and possible future of digital humanities. We develop this approach in the context of these new techniques of knowledge-presentation, new infrastructures for knowledge work, and new formations around human capacities to work with complex and large data sets. Strong claims are often made about the potential for replacing aspects of traditionally humanities work undertaken by human labour alone through machine-learning techniques. Here, through a critical examination of new epistemologies and machine-generated data ontologies, for example, we examine the possibility of methods for a *critical digital humanities* in relation to new machine-learning techniques, together with how machine learning might be repurposed *for* a critical project within DH scholarship.

Towards a critique of machine learning: critical digital humanities and AI

David M. Berry

In this paper I investigate the claims of computational models and practices drawn from the field of artificial intelligence and more particularly machine learning. I do this to explore the extent to which machine learning raises important questions for our notions of being human, but also, relatedly the concept of civil society and democracy as distilled through notions of hermeneutic practice. That is, that in the 21st century we are seeing the creation of specific formations which threaten historical notions of humanities research and thinking. They represent new modes of knowing and thinking driven by these new forms of computation such as machine learning and Big Data, and which will have implications for the capacity to develop and use social and human faculties.

It is certainly the case that through the innovative assembling and organisation of scale technologies together with human actors new cognitive forms are under construction and experimentation. This paper develops a speculative and theoretical discussion of possible future directions driven by what we call Informating or Augmenting technologies in the digital humanities. In this paper, the notion of a digital humanities is linked to the social, cultural, economic and political questions of a recontextualisation and social re-embedding of digital technologies within a social field. Indeed, exploring the digital humanities through a critical lens I seek to understand how different disciplinary specialisms are newly refracted not just by their interaction, but also by the common denominator and limitations of computation. That is, how the constellation of concepts that are used within a disciplinary context are challenged and transformed within a computational frame.

Indeed, this raises theoretical and methodological questions, for example, digital humanities is keen develop tools to explore the new techniques such as machine learning for the field. This calls for a critical response, and there has already been some valuable work undertaken in this area, such as Alan Liu's work on critical infrastructure studies, but here I explore how a critical digital humanities can offer a way of thinking about the theoretical and empirical approach to massive-scale technologies. In this paper I argue that digital humanities should not only map these challenges but also propose new ways of reconfiguring research and teaching to safeguard critical and rational thought in a digital age. First, I turn my attention to research infrastructure and how critical approaches can contribute to and offer methods for contesting ML. I argue that research infrastructures provide the technical a priori for the support of and conditions of possibility for digital humanities projects, but in a machine-learning paradigm different techniques and critical methods will be required to make sense of their use. Secondly, in relation to data, we might consider the more general implications of datafication not just within the general problem of big-data, but in terms of the specific issues raised by machine learning in the generation, processing and automated classification of data–especially where the metadata becomes nonhuman-readable.

This links to my final question about how visibility is made problematic when mediated through computational systems. The question is also linked to *who* and *what* is made visible in these kinds of machine-learning systems, especially where as Feminist theorists have shown, visibility itself can be a gendered concept and practice, as demonstrated in the historical invisibility of women in the public sphere, for example (see Benhabib, 1992). Finally, this paper will explore how to embed the capacity for reflection and thought into a critically-oriented digital humanities and thus to move to a new mode of experience, a two dimensional experience responsive to the potentialities of people and things intensified by the advances in machine-learning capacities. In other words, the reconfiguring of quantification practices and instrumental processes away from domination (Adorno, Horkheimer, Marcuse) and control (Habermas), instead towards reflexivity, critique and democratic practices. As Galloway argues, "as humanist scholars in the liberal arts, are we outgunned and outclassed by capital? Indeed we are–now more than ever. Yet as humanists we have access to something more important.... continue to pursue the very questions that technoscience has always bungled, beholden as it is to specific ideological and industrial mandates" (Galloway, 2014: 128). I argue that specific intervention points within the materialisation of this ML a priori, such as in design processes, can be explored to contest machine-learning techniques that serve to instrumentalise humanities approaches.

Digital humanities has the technical skills and cultural capital to make a real difference in how these machine-learning projects are developed, the ways in which instrumental logics are embedded within them and interventions made possible. For example, digital humanities through its already strong advocacy of open access, could push for and defend open source, open standards and copyleft licenses for technical components and software, opening up and documenting new techniques for machine-learning by humanists for humanists–but this could also be the opening up of the complexity of the black box of ML systems. The ways in which these aspects interrelate in terms of the ML "space of work" is hugely important, that is, the functional capacity of a machine-learning system is crucial, in as much as the range of humanities work may be adversely effected or inhibited by the shape of a machine-learning infrastructural system. I argue that these are urgent questions, with the recent turn towards what has come to be called "platformisation", that is the construction of a single digital system that acts as a technical monopoly within a particular sector, and it is certainly the case that the implications of machine-learning infrastructures and their black-boxed techniques for sorting, classification and ordering large amounts of data needs constant vigilance from digital humanists.

Augmenting and automating human and machine attention in the (digital) humanities

M. Beatrice Fazi

Attention denotes the cognitive process of selecting and focusing upon certain aspects of information whilst ignoring others. In recent years, it has been argued that this special state of percipient awareness is undergoing a profound transformation, due to the increasing intertwining of digital devices and everyday cognitive tasks (see Carr, 2010; Gazzaley and Rosen, 2016). Social media, phone apps, design interfaces, smart devices: the industry markets these technologies as helpful assistants that will free us from the chore of identifying, selecting and retaining relevant information, thus allowing us to dedicate our time, and our mental efforts, to other things. In addition, digital software and hardware are equally used to tune senses and to maintain motivation. Whilst cognitive cognates such as memory and intelligence are of course also targeted, it is the capacity to pay attention that seems primarily to be called into question here. In an attention economy, attention is believed to be a scarce commodity. The assumption is that, with the current information overload, digital machines are instruments able to outsource decisions regarding what to prioritise, what to select and what to discard in the data-deluge. Whilst much concern in the 20th century focused on the question "Can a machine think?", and Artificial Intelligence labs were devoted to answering this question, in the 21st century the central question seems to be "How do we think with machines, and how do we get machines to do much of our thinking for us?"

The exteriorisation of cognitive faculties such as the capacity for attention does, however, come at a price. Studies in neuroscience and neuropsychology, drawing from theories of neuronal plasticity, show that our brain is being rewired in favour of new cognitive skills, and to the detriment of older but cherished abilities, such as the capacity to read a novel from cover to cover. Evidence of this deterioration of human attention comes from science, yet everyday anecdotal confirmations also come from educators and parents, who report of children who cannot focus, and of students who are distracted and cannot complete their assignments. Relevantly, N. Katherine Hayles (2007) has described this situation in terms of a generational cognitive shift.

The humanities, due to the fact that they are largely based around texts, have often elaborated and developed concerns about human attention under the rubric of debates as to what counts as reading. Within the digital humanities, more specifically, it has been stressed that, whilst humans are very good at "close reading" (i.e. the careful, attentive and sustained inspection of a text), computing machines allow us to consider a broader picture. Franco Moretti (2013), amongst others, has called this condition "distant reading". These debates have opened up considerations about the possibility of an "algorithmic criticism" (Ramsay, 2011), as well as reflections on the importance of the hermeneutic faculties of human beings (Berry, 2012; Stiegler, 2010 and 2016). In this paper, I depart from these discussions within the digital humanities and then move to argue how new understandings of human attention might emerge in conjunction with possible conceptions of what I call "machine attention". I will map these possible conceptions of machine attention in relation to increasingly popular Artificial Intelligence techniques known as machine learning. More specifically, I will consider how machine-learning programs might be said or seen to pay attention to data-stimuli: they detect some information and discard some others, forming and dissolving patterns, in order to shape and sharpen their cognitive outcomes based on these selections. I will then emphasise the relevance of these modes of machine attention for the way in which we can understand what human attention might become after the computational turn in the humanities.

The question of what is happening to human attention is an important and pressing one for the humanities. It is always difficult to define what the humanities are, or where they begin and end. However, surely few would object that the humanities are the locus of "deep" attention: humanities disciplines prioritise textual analysis, where the process of knowing is intimately connected to those of making sense, interpreting, and of giving meaning. These are epistemic processes that start and end with the cognitive exercise of attention. The pedagogical issue of what happens to students if they have lost (or never gained) the capacity to focus is also a question upon which the future of humanities disciplines, and humanities departments, seems to be predicated. In this paper I will address these issues, by considering the intermeshing of human and machine modes of attention, whilst also arguing that our engagement with automated forms of attention (as well as of other automated and augmented cognitive processes) should involve a commitment to re-defining and enlarging the prospect of what computational mechanisms are, and what rule-based, computational cognitive processes might amount to.

'The new spirit of automation': the changing discourse of automation anxiety

Caroline Bassett, Ben Roberts and Jack Pay

From self-driving cars, through high-frequency trading to military drones and organised swarms of shelf-stacking robots, our era is marked by rising automation and a new fascination with the likely social, cultural, and economic impacts of this computationally driven transformation. This paper will explore innovative methods by which the humanities might address contemporary *automation anxiety*. The wider topic of automation is a pressing subject with various existing academic responses such as Frey and Osborne's work on automation and the future of employment (2013). The focus of this paper is to address, as a topic in its own right, the cultural and social anxiety generated by these new forms of computational automation. What new research methods can the humanities use to map and understand automation anxiety around opaque computational decision making? What digital tools can be brought to bear on the diverse types of online public culture in which this anxiety is expressed?

Automation anxiety is evident in a plethora of popular contemporary accounts, public debates and political interventions. Tyler Cowen's *Average is Over* depicts a dystopian future in which the job market is divided between a highly educated and skilled elite capable of harnessing automation for personal wealth creation and a wider mass who are consigned to low paid work. Other accounts see in this new wave of computerisation the potential for a productive redefinition of the relationship with work. Futurists Martin Ford in *Rise of the Robots* (2015) and Jerry Kaplan in *Humans Need Not Apply* (2015) propose to respond to the automation of work through the creation of a universal income. In a more radical version of this thesis, postcapitalism, as charted by Paul Mason, posits automation as the basis of a technologically-driven, non-market successor to capitalism. Another type of anxiety arises out of the increasing use of computerisation in law enforcement and military action. Here there is an automation anxiety that the current wave of military drones will evolve into fully autonomous killing machines, with software systems governing decisions about life and death. In July 2015 over 3000 robotics and artificial intelligence researchers and over 17000 other academics and interested parties (including Stephen Hawking, Elon Musk and Noam Chomsky) signed an open letter, published on the Future of Life website and widely disseminated in the global media, calling for a global ban on "offensive autonomous weapons beyond meaningful human control." There is also a more general anxiety which asks what happens to human life when so many tasks are automated away. Nicholas Carr's *The Glass Cage: Where Automation is Taking Us* (2014) suggests that automation is a threat to humanity itself—as we delegate tasks to computational tools, human cognitive capacities atrophy, understanding weakens, and the power of human reasoning is undermined.

This paper places contemporary automation anxiety in the context of historical debates about automation. It examines methods that might be used to analyse changing social attitudes to automation and computation between the 1960s and the present. Automation was a controversial topic in both Britain and the United States in the 1960s. In 1964 defence automation specialist Sir Leon Bagrit gave the public BBC Reith lectures on the topic. In the same year, President Lyndon B. Johnson set up the National Commission on Technology, Automation, and Economic Progress. Then, as now, there were concerns about automation and the future of employment. Then, as now, there were utopian imaginings of the future social benefits of automation. Nevertheless the hypothesis here would be that there are important differences between the two eras and that we can learn from changing attitudes to automation and computation. Among other things, analysis of changing attitudes to automation might illuminate different historical perspectives on: the end(s) of work; the relationship between labour and the domestic sphere; the role of computation in society.

The paper takes inspiration, but not theoretical orientation, from Boltanski and Chiapello's *The New Spirit of Capitalism* which used textual analysis of management literature from the 1960s and 1990s to argue for a fundamental shift in what they call the 'spirit of capitalism', i.e. the way in which capitalism justifies itself. In a similar vein we analysed key grey literatures (policy, commercial reports, academic and government papers) on automation from the 1960s and present in order to understand the changing discourse around automation. A key concern was to use digital humanities tools which provide different scales of analysis and new perspectives. We did this both to generate new understandings of automation anxiety across time and to investigate ways in which digital humanities and media archaeological approaches intersect.

The "new spirit" of capitalism which has emerged between the 1960s and the present day consists in a highly decentralised networked form of capitalism, characterised by "flatter" organisational hierarchy, much greater autonomy within firms for both individuals and teams, lower job security and the proliferation of temporary contracts and outsourcing. Boltanski and Chiapello use their sociological analysis of management literature to support a more speculative, philosophical account of capitalism and its critique, notably seeing the contemporary spirit of capitalism as incorporating the critiques that were made of capitalism in the 1960s and particularly around May 1968.

Similarly this paper argues that automation controversies could be a springboard to more general debates about the changing relationship between computation and society. The central premise here is that there is a much to be discovered from *attitudes to* automation and the justification of computation tools as there is from the specific technological forms and implementations of automation.

Bibliography

Bagrit, L. (1965). *The Age of Automation*. The BBC Reith Lectures 1964. London: Weidenfeld and Nicholson.

Benhabib, S. (1992). *Situating the Self: Gender, Community and Postmodernism in Contemporary Ethics*. London: Routledge.

Berry, D. M. (2011). *The Philosophy of Software: Code and Mediation in the Digital Age*. London: Palgrave Macmillan

Berry, D. M. (2012). *Understanding Digital Humanities*. Basingstoke: Palgrave

Berry, D. M. (2014). *Critical Theory and the Digital*. New York: Bloomsbury

Berry, D. M. and Fagerjord, A. (2017). *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity.

Boltanski, L. and Chiapello, E. (2007). *The New Spirit of Capitalism*. London: Verso

Carr, N. (2010). *The Shallows: What the Internet is Doing to Our Brains*. New York: W. W. Norton

Carr, N. G. (2014). *The Glass Cage: Automation and Us*. New York: Norton.

Cowen, T. (2013). *Average Is Over: Powering America Beyond the Age of the Great Stagnation*. New York: Dutton.

Ford, M. (2015). *The Rise of the Robots: Technology and the Threat of Mass Unemployment*. London: Oneworld Publications.

Frey, C. B. and Osborne, M. A. (2013). *The Future of Employment: How Susceptible Are Jobs to Computerisation*. Oxford Martin Working Papers.

Kaplan, J. (2015). *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*. New Haven, CT: Yale University Press.

Galloway, A. (2014). “The cybernetic hypothesis.” *differences*, 25(1): 107-31.

Gazzaley, A. and Rosen, L. D. (2016). *The Distracted Mind: Ancient Brains in a High-Tech World*. Cambridge, MA: The MIT Press.

Gold, M. K. (2012). “The digital humanities moment.” In M. K. Gold (ed) *Debates in the Digital Humanities*. Minneapolis, MI: University of Minnesota Press, pp. ix-xvi.

Hayles, N. K. (2007). "Hyper and deep attention: the generational divide in cognitive modes." *Profession 2007*, pp. 187-99.

Mason, P. (2015). *Postcapitalism: A Guide to Our Future*. London: Allen Lane.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Ramsay, S. (2011). *Reading Machines*: *Toward an Algorithmic Criticism*. Champaign, IL: University of Illinois Press.

Stiegler, B. (2010). *Taking Care of Youth and the Generations*. Translated by S. Barker. Stanford, CA: Stanford University Press.

Stiegler, B. (2016). *Automatic Society*. *Volume 1: The Future of Work*. Cambridge: Polity.

Archaeologies of Reading: Modeling and Recreating the Annotation Practices of Gabriel Harvey, John Dee, Jacques Derrida, and the Winthrop Family

Jean Bauer
jabauer@princeton.edu
Princeton University, United States of America

Earle Havens
ehavens2@jhu.edu
Johns Hopkins University, United States of America

Matthew Symonds
m.symonds@ucl.ac.uk
University College London, United Kingdom

Anthony Grafton
grafton@exchange.princeton.edu
Princeton University, United States of America

Jennifer Rampling
rampling@exchange.princeton.edu
Princeton University, United States of America

Katie Chenoweth
kac3@exchange.princeton.edu
Princeton University, United States of America

Christian Flow
cflow@exchange.princeton.edu
Princeton University, United States of America

This panel brings together three projects that explore and analyze annotated texts in digital environments: The *Archeology of Reading in Early Modern Europe*, The Winthrop Family on the Page, and Derrida’s Margins: . All three projects seek to give users the experience of how another person read their books (see Notes). While each is fascinating in its own right, collectively these projects span centuries and provide a powerful and instructive lens on data modeling, inter-institutional collaboration, and interoperability.

The connections between these projects, both on a scholarly and an institutional level provide a powerful case study in creating distinct but interoperable digital humanities projects and resources. Both Winthrop and Derrida’s Margins are being built in the Center for Digital Humanities at Princeton University, while AOR is a collaboration between Johns Hopkins, University College London, and Princeton. AOR and the Winthrop project share a co-PI, Anthony Grafton, and two books owned and heavily annotated by John Dee are in the Winthrop library and thus came to New England. However, while the AOR project looks at a single renaissance annotator (Gabriel Harvey in Phase I– the project is heavily indebted to the methods and questions outlined in Jardine and Grafton, 1990–, and John Dee in Phase II), the Winthrop project looks at generators of annotators (men and women) making it as much a prosopographical project as a study of annotation. Like AOR, Derrida’s Margins focuses on a single scholar’s interaction with his library, but Jacques Derrida’s annotation practices differ dramatically from Harvey or Dee, and French copyright law poses serious challenges to representing the book pages online.

All three projects make use of the International Image Interoperability Framework, but differ on the rest of their underlying data structures. AOR uses a custom XML schema to encode a critical edition of the annotations. Winthrop will use a custom prosopographical-bibliographical relational database, implemented in Django. While most of these relationships will be attested through annotation, the system will also record other connections (including purchasing records, books referenced in letters, etc). The annotation data model will most likely be created using a graph database, given the multiple types of annotations and the ways they can be expressed. Derrida's Margins will be another custom relational database, built in Django, but will be fully bi-lingual, allowing users to search and browse in English and French. The Winthrop and Derrida database will also be exposed as Linked Open Data.

All three projects are committed to interoperability, with a key goal being the ability to search across the three projects once they are complete. The projects are in close contact the IIIF Editors and the W3C Web Annotation Working Group to ensure adherence to and input on best practices and emerging standards.

The Archaeology of Reading in Early Modern Europe

Earle Havens, Matthew Symonds,
Anthony Grafton

The Archaeology of Reading in Early Modern Europe (AOR), is an international collaboration among the Sheridan Libraries at Johns Hopkins University, the Centre for Editing Lives and Letters (CELL) at UCL, and the Center for Digital Humanities at Princeton University Library, with funding from the Andrew W. Mellon Foundation. It is also an interdisciplinary collaboration between historians, librarians, software engineers and data scientists.

While the body of scholarship on the history of early modern reading practices has burgeoned during the past several decades—guided in large part by the initial scholarly work of our project partners, the late Professor Lisa Jardine and Professor Anthony Grafton—as a collective body of knowledge the history of reading has nonetheless remained limited to isolated, partial, and impressionistic studies of single texts read by single annotators.

As researchers, we conduct this work in the conspicuous absence of comparative evidence of the larger range of early modern historical reading practices, strategies, and agendas. Scholars also find it physically impossible to effectively penetrate the dynamic array of information preserved in annotated books for the purpose of systematic analysis owing to their sheer density of content, in many instances, relative to the original texts on which the annotations comment.

During the original planning workshop on annotated books that led to the formation of the project, leading scholars, librarians, curators, and technologists agreed collectively that these constraints imposed upon the study of early modern annotated texts in their original analog form could only be overcome satisfactorily when annotations are treated as data sets that can be mined and analyzed effectively in more versatile, enriched, and readily searchable digital forms.

The AOR team has elected to focus on a distinct and roughly contemporary dyad of clearly identified early modern readers: Gabriel Harvey in Phase 1 (2014–2016) and John Dee in Phase 2 (2016–2018). While the identities of a large majority of early modern annotators remains unknown in extant collections, this focus on known readers will enable the project team to analyze and more precisely situate the processes of reading and annotation within their respective historical contexts.

In order to make these sources more accessible to analysis, the AOR technology team – based at Johns Hopkins University's Digital Research and Curation Center – is working closely with the International Image Interoperability Framework (IIIF) protocol and community to develop features and use cases that will enhance AOR within this larger framework.

IIIF features a set of protocols, application programming interfaces (APIs), and shared technologies for the presentation of web-based images. In the case of AOR, these images are digital surrogates of rare book materials containing manuscript annotations by Gabriel Harvey and John Dee, which users interface through AOR's adapted version of the Mirador (version 2.0) viewer. The technical infrastructure for AOR includes a data archive, an image server, a IIIF image service, a corresponding IIIF presentation service, and the IIIF-compliant Mirador viewer.

The data archive for this project provides the framework for long-term access to, and preservation of, all project content: an important contribution to Digital Humanities more generally, insofar as this issue has not yet been unaddressed within IIIF. The technology team has defined an archival data model that can be mapped to other data models for data access and presentation over time. Another layer of the infrastructure that has been developed in Phase 1 of AOR consists of an image server that accesses content from within the data archive. Currently, the team is utilizing a commercial FSI image server, though comparable image server resources (e.g. djatoka) may also be used.

While AOR has been deployed using the versatile Mirador 2.0 API, the AOR technology team has developed IIIF endpoints for any available image service and presentation service, so long as they can be accessed by an IIIF-compliant image viewer. All AOR data can be accessed through our Mirador2 IIIF-complaint viewer, which has been specifically enhanced to meet the use requirements identified by the team for both current and future users. Over the course of AOR Phase 1, the technology and scholarly teams have worked together closely to define and implement a set of use cases related to image viewing and manipulation, transcription viewing, and dynamic, query-building search capabilities.

The current AOR viewer is the culmination of this iterative development process, which presents users with a wide-ranging set of functionalities aimed at enabling new forms of research on the history of reading practices. One new method is informed by the use of insights gleaned from the data generated throughout Phase 1 to formulate new research questions for the humanities team to investigate.

By creating a corpus of important and representative annotated texts with searchable transcriptions and translations, we can begin to compare and fully analyze early modern reading, and place that mass of research material within a broader historical context. In so doing, we could also approach—not in isolation but as a dynamic, internally and institutionally complimentary, research team—the traditionally subjective study of reading in a demonstrably empirical, comparative, and systematic way.

All visible interventions by readers in the books – marginal notes, underlining, marks and symbols, et cetera – have been marked up according to a non-TEI XML schema developed by humanities and technology teams working in close cooperation. The Phase 1 corpus of thirteen books owned and annotated by Gabriel Harvey has generated 3,355 XML files, each representing a single page of an annotated book. The data contained within these XML files has allowed us to map the language Harvey used in his own annotations and highlighted in the original printed texts, to find interesting correlations, and to use those correlations to form new investigations into the history of reading, Harvey's own biography, and the wider history of ideas in the early modern period.

Derrida's Margins: Annotations from the Personal Library of Jacques Derrida

Katie Chenoweth

This paper will give an overview of Phase One of the Derrida's Margins project currently underway at the Center for Digital Humanities at Princeton University (CDH).

Derrida's Margins is a longterm project that aims to create a website and online research tool for annotations from the Library of Jacques Derrida, an archival collection housed at Princeton University Library's Rare Books and Special Collections that was first opened to the public in March 2016. Phase One of Derrida's Margins focuses on annotations related to Derrida's landmark 1967 work De la grammatologie, first translated into English in 1976 as Of Grammatology (hereafter OG, see Spivak's 2016 translation. This corpus will serve as a pilot data set for future work, allowing us to establish protocols, workflow, and a relational database model.

Jacques Derrida is one of the major figures of twentieth-century thought, and his personal library represents a major intellectual archive. The Derrida Library, consisting of about 19,000 published books and other materials, represents a lifetime of reading. But for Derrida, the act of reading was not a passive process: he engaged — even grappled — with what he read, covering pages with notes and cross-references, inserting other handwritten materials, quoting and adapting what he read into what he wrote. As Derrida himself said in an interview later in his life, his books bear "traces of the violence of pencil strokes, exclamation points, arrows and underlining" (for details related to Princeton's acquisition of the Derrida Library, see the article posted by Princeton University Libraries, 2015).

It was in OG that Derrida first articulated a new style of critical reading, which would become the foundation of the philosophy of "deconstruction." Our online research tool will enable scholars to study the development of this philosophy in an unprecedented way by providing comprehensive digital access to the material annotations, marginalia, bookmarks, and other notes from Derrida's library that correspond to each quotation and citation in OG. Beyond making Derrida's annotations available digitally for the first time, this project seeks to enable researchers to understand the relationship between Derrida's published writing and his reading practices. As we move beyond OG in future phases, the website will also allow researchers to gain new insights into Derrida's library and his published writing as networks of texts, citations, and annotations.

OG was chosen as the pilot text not only because it is a foundational text for deconstructive reading, but for two additional reasons: 1) it is among the most widely read of Derrida's works and, according to Google Scholar, by far the most frequently cited, constituting more than 10% of the total citations for Derrida as of November 1, 2016; 2) 2017 will mark the fiftieth anniversary of the initial publication OG, and a number of scholarly conferences and events are scheduled to discuss the text and its legacy.

In Phase One, we began by identifying all instances of citation (quotations, references, footnotes, etc.) in OG that lead us back to books and other reading materials from Derrida's library. Each cited work from OG is entered into a Zotero library, which will eventually form the basis of a public bibliography for OG. We then located these references in Derrida's copy (or, often, multiple copies—different editions, translations, etc.) of each work. Each instance of citation is given a tag in the Zotero library that indicates the source page from OG, the type of citation, the page number(s) in the cited text, whether or not the volume is present in Derrida's Library at Princeton, and the presence or absence of annotation (here construed as a mark of any kind, verbal or non-verbal). Next, we will transcribe all marginal annotations and other markings, limiting ourselves to tagged pages, i.e., those that are explicitly referenced by Derrida in OG. Digital images and transcriptions of these annotated pages will form the basis of the website, allowing users to go "behind the scenes" of Derrida's reading practices.

Given that Derrida's work is widely read in the United States, France, and around the world by scholars from numerous disciplines in the humanities and social sciences, we anticipate that the audience for this project will be broad and international. The more targeted audience would be specialists of Derrida and deconstruction, as well as researchers in philosophy, literary studies, and intellectual history.

One major risk this project faces is the question of French copyright restrictions. Due to the fact that the majority of books owned by Jacques Derrida are copyrighted works whose copyright is still in force, there is a question regarding the legality of posting digital images of sections of those books on our website. Despite the fact that the copies of these works in JD's library have been annotated by the philosopher's hand, the works themselves are the intellectual property of their respective publishing houses in France. Pending clarification by the Office of the University Counsel regarding the amount of risk we are at liberty to take in this respect, we will make a decision as to the amount of text from those works that it is acceptable for us to display on our public website.

This project is intended as the first phase in a long-term project. Team members in Phase One will create a manual detailing their work process and workflow to facilitate and standardize the project going forward. At the end of our pilot phase with the CDH, we will have the following

The main outcomes and deliverables of this project are the following:

- Bibliography of *De la grammatologie,* made available as public Zotero library
- Digitization of all relevant pages from the Derrida library
- Customized grayed-out images, if necessary
- Annotations transcribed, translated, and tagged
- Custom designed relational database to record and analyze Derrida's annotation practices and the anchor text they reference
- Bi-lingual (French and English) Web portal created in Django to allow users to search and browse the annotations
- Transcriber's manual

The Winthrop Family on the Page

Anthony Grafton, Jennifer Rampling, Christian Flow

"The Winthrop Family on the Page" will employ an extensive database of bibliographic information, biographical data, and marginal annotations to allow digital exploration of early modern lives and learned practices. Centering on the surviving library holdings of the storied Winthrop family, whose representatives included such prominent figures as John Winthrop, a founder and governor of the Massachusetts Bay colony, the final product will be a web platform affording a dynamic sense of how colonial readers interacted with texts and fellow readers. Users will be encouraged to follow the story along multiple axes, both diachronically—as family members communicated across decades, even centuries, in the margins of their shared books—and synchronically, as single readers followed references from one text to the other, leaving a bread-crumb trail of notation along the way. The experience will be further enriched by inclusion of information on how the Winthrops' books surface in other historical sources they left us, including journals and correspondence, ensuring once again vivid access not just to the texts, but to the people who read them.

The bulk of the project's database will consist in entries for some 300 books formerly owned by the Winthrops and currently held at the New York Society Library. Aside from bibliographical information for all of these books, the database will include a second module treating the marginalia they reveal. Here, the project will log the location of all manuscript notes in the collection; in addition, for a select, smaller subset of 50-60 books, high-quality digital images will allow users to view the marginal material directly. Several representative notes from the same sub-set will be translated and transcribed along with relevant anchor text, and all of the collected data will be encoded for search. Users will be able to come to the collection with queries about everything from annotations of a particular era, to those of a particular person, to those presenting a specific sign (e.g., the manicule) or handling a specific theme.

The defining characteristic of the project is its deft exhibition of a very particular source-base. The rich surviving Winthrop holdings, coupled with the opportunities for collection and display afforded by a digital platform, offer a rare opportunity to excavate and reassemble an Early Modern book collection, giving a concrete frame to the "intellectual space" within which the family lived and thought. Still more exciting is the chance to fill that space with dialogue: the project's careful curation, transcription, and (when necessary) translation of the marginalia that the Winthrops recorded as they read allows users to follow their intellectual journey between books and between generations. Those generations spanned not just time but space: because the Winthrops were a colonial family, their New England library had its roots and its first readers in Europe. Users of the collection are therefore positioned to pose and answer questions about how Early Modern knowledge made the transition from the Old to the New World. How, we wonder, might our understanding of the Salem Witch Trials be affected when we take into account the fact that Wait Still Winthrop, the Chief Magistrate of Massachusetts, could well have perused the marginal annotations of his European ancestor, Adam Winthrop, in the witchcraft section of the family library? How did reading and annotation practices originally cultivated in England filter into the learned arsenal of later colonial readers?

Queries like these are made actionable by the labors of a highly-skilled team: specialties in book history and annotation, in ecclesiastical history, in alchemical practice, and in the classical tradition will be amply represented. The project itself, which will offer direct access to an array of understudied material, is meant to have wide interdisciplinary appeal. Situated at the nexus of digital humanities and history of the book, straddling Europe and the colonies, the domains of intellectual history and learned practice, it will resonate with historians of Europe, America, and the Atlantic World; book historians; librarians; historians of science, medicine, and religion; and users outside the Academy. And in future years the appeal will only grow more broad-based, for project members hope to build outward from the corpus on which the 2016-17 work is founded. The more material at hand, the richer the web of connections users will be able to make between books and readers. There are several possible modes of expansion: (1) full imaging, transcription and translation of all books in the NYSL corpus (that is, including those beyond the subset currently targeted for the pilot-phase of the project); (2) inclusion of further Winthrop books housed at other major repositories, including items from the eighteenth and nineteenth centuries (when the Winthrops added hundreds of new books to the library, including a collection of almanacs now housed at the Houghton Library, and the bulk of the collection of Winthrop books at Allegheny College); (3) development into a broader study of Colonial reading practices and knowledge networks, involving other family collections of annotated books, such as the Mather Family Library, now held at the American Antiquarian Society; the library of Thomas Prince, now at the Boston Public Library; and the library of James Logan, now at the Library Company of Philadelphia.

Notes

1. There are a number of projects dedicated to recording annotations or marginalia, notably the crowd sourced Annotated Books Online and Book Traces. The three projects discussed in this panel proposal differ in that they are specifically grouped by annotator and want to answer specific questions about the worldview and read patterns of known annotators. They are also designed to facilitate more complex queries. These projects are also addressing a different question from other annotation projects such as Annotation Studio which are designed for modern readers to annotate digital texts.

Bibliography

Derrida, J. (1967), *De la grammatologie* (Paris: Editions du Seuil).

Derrida, J., (2016) *Of Grammatology*, trans. Gayatari Chakravorty Spivak (Baltimore: Johns Hopkins University Press).

Jardine, L., and Grafton, A. (1990)"'Studied for Action': How Gabriel Harvey Read His Livy," *Past & Present*, no. 129: 30–78.

Princeton University Libraries (2015). Princeton University Library Acquires Jacques Derrida's Personal Library. 31 March, http://library.princeton.edu/news/2015-03-31/princeton-university-library-acquires-jacques-derridas-personal-library

Scaling up Arts and Humanities: The DARIAH Approach to Data and Services

Aurélien Berra
aurelien.berra@u-paris10.fr
Université Paris-Ouest Nanterre La Defense, France

Matej Durco
matej.durco@oeaw.ac.at
Austrian Center for Digital Humanities, ÖAW, Austria

Chad Gaffield
gaffield@uottawa.ca
University of Ottawa, Canada

Nicolas Larrousse
nicolas.larrousse@huma-num.fr
Centre National de la Recherche Sienctifique, France

Paulin Ribbe
paulin.ribbe@huma-num.fr
Centre National de la Recherche Sienctifique, France

Mike Priddy
DANS - KNAW, Netherlands
mike.priddy@dans.knaw.nl

Carsten Thiel
thiel@sub.uni-goettingen.de
SUB Göttingen, Germany

Introduction

The panel contributions will introduce and discuss several central aspects of DARIAH-EU aimed to support emerging practices in the Digital Humanities, to disseminate consolidated practices and to provide tools for navigation through the rich and changing landscape of the productions of research communities, be they research data, tools, methods or other assets.

Our focus point is how to boost research on a number of different levels: ensuring the sustainability of the infrastructure, its "scaling up" to enable better access to research data, community-building, linking teaching activities and weaving the network of affiliated initiatives.

Panel Overview

Chad Gaffield (Moderator)

Panel participants as listed below will introduce their topic for 5 minutes each, followed by a general discussion.

It has been much debated in the Digital Humanities communities whether scaling up – in terms of shared resources, formation of team science, cross-domain research – is at all appropriate for the way knowledge is produced in the various humanistic disciplines. Instead of tackling again this general problem, this panel takes a much more practical approach. There can be no doubt that research practices in the humanities and social sciences are profoundly changing (Wouters et al., 2013). The turn to digital methods happens in a relationship of mutual dependency and co-evolution with the infrastructures traditionally feeding humanities research: collections, libraries and archives. This development is most visible in large-scale research infrastructures such as CLARIN and DARIAH, and it is acknowledged by their respective funding.

This panel presents the efforts made within the DARIAH (Digital Research Infrastructure for the Arts and Humanities) community in its mission to further develop the Arts and Humanities research infrastructure, now that the Digital Humanities have long left the incubation stage or the phase of early adopters. The field is still growing but also starting to differentiate and specialise. The “scaling up” of the humanities refers to the growth of the community – as a social, cognitive and technological network –, the increasing availability of Big Data, the ongoing penetration of standards and the use of consolidated tools. While the mission is quite clear, its implementation is a far more complex process.

We will focus on the relevance of the **community** and **sustainability** as driving factors in building the research infrastructure**.** The contributions report about lessons learned, examine achievements and open questions, and aim to engage with the audience about their expectations, experiences and visions of what a research infrastructure (RI) for the Arts and Humanities should look like.

These two aspects are not arbitrary. Community-building responds to the motto of DARIAH: *services developed for researchers by researchers* and is relevant to formulate needs, to find appropriate technical solutions but also to encourage continuous (re)use by means of education and training. DARIAH itself is a pan-European infrastructure, driven by grass-rooted needs of researchers, and governed by national representatives, entailing processes of collaborative evaluation. In addition to community-building within the RI, DARIAH has always aimed to expand and strengthen the external cooperations with a broader ecosystem involving both affiliated RI projects (EHRI, CENDARI, PARTHENOS) and sister initiatives like CLARIN, as well as with research infrastructures that specialize in providing basic services (eduGain, EGI or EUDAT). Research Infrastructures address both their users and producers with their immediate needs. At the same time they have to be stable, reliable and of good quality, which needs to be addressed continuously throughout its life. Another primary concern of DARIAH is the focus on reuse of research data and ensuring their longevity. This is where the sustainability of software and services comes into play, ensuring the continued availability of the outputs from shorter-lived projects in the long run.

Building a durable infrastructure

Nicolas Larrousse and Paulin Ribbe

During the last decades, digital made his way into the Arts and Humanities research world at every stage of research projects. As a consequence, the way researchers work today has considerably evolved. They can no longer work isolated and their work requires more and more sophisticated tools to deal with research material. There is a need to scale up and build infrastructures dedicated to Arts and Humanities. DARIAH is one possible answer to these new requirements.

DARIAH is organized as an European framework for research infrastructures based upon a consortium of states which agree to support the infrastructure in the long term. It was designed to establish sustainable digital services and develop common practices. Such objectives reflect a long-term vision which would not fit in a “classical” project funded for three years.

Within the consortium, ongoing efforts help to concretely build the infrastructure and to better answer those fundamental questions: what is an infrastructure for Arts and Humanities and how can we build it on the European level?

DARIAH contributions and the reference architecture they inhabit

Mike Priddy (Co-authors: Francesca Morselli, Lisa de Leeuw, Andrea Scharnhorst)

DARIAH constitutes itself as a pan-European infrastructure and relies on national contributions from its member states. 90% of these contributions are in-kind and represent the research outputs that each member country contributes to the infrastructure, showing the richness and diversity of the European research landscape. DARIAH is creating a way to collect, disseminate, review, assess and evaluate the contributions made to the infrastructure and community.

As researchers in the Arts and Humanities familiarise with the creation, management, storage and reuse of data, they need a technical infrastructure that can support such activities. On the other hand, a mere technical support seems insufficient to sustain their research activities, as research communities play a crucial role in the development of the infrastructure by developing processes and offering services and resources. An approach which supports the knowledge creation process as well as an agreement on standardised processes (e.g. data exchange) is therefore highly needed.

With a reference architecture we aim to create a common language among the European member countries per individual type of contribution to describe and communicate DARIAH contributions, as well as provide a reference for other projects and infrastructures in the Arts and Humanities.

Re-usability of tools and services

Carsten Thiel

The software tools and solutions powering the services provided by infrastructures are subject to their individual life cycles. While most projects realised through research grants focus on the initial phases – from inception and initial design, through user testing and design revisions, to beta releases –, delivering a "final" product by the project's end remains the primary objective. But the software's life does not end there. As researchers use the tools, new needs and requirements can arise that need to be addressed and implemented. At the same time the IT world evolves and technologies change. This leads to an increased need for adopting existing services and their underlying software for continued usability. Finding resources to sustain the services and to ensure their availability and accessibility requires a change already to the initial objectives in the development phase (see Doorn, Aerts, Lusher 2016). Designing software from the get-go with re-usability in mind ensures the long-term benefit of enabling future contributions and adoptions needed by the community as a whole.

Teaching: where and how?

Aurélien Berra

DARIAH's "Research and Education" programme has contributed or planned several services which emphasize how data are central in building up a sense of community and deliberate, open, collaborative research strategies.
Three projects will be discussed. The Digital Humanities Course Registry is a collaborative database aiming to provide up-to-date information on European courses to students looking for a programme and to teachers and researchers interested in evaluating the state of the field at a national or international level. The initiatives comprised in the "Integration of Training Material" project offer a framework for sharing and preserving Digital Humanities teaching materials (through the user-centred design of #dariahTeach, an open platform for extensible and translatable contents,) as well as a philological toolbox based on the principle of re-use and adaption (Biblissima's BaOBab, a catalogue of online tools, guidelines and tutorials focused on manuscripts traditions,). The Master Classes scheduled for 2017 in Maynooth (Public Humanities), Florence (Evaluating Digital Scholarship), Berlin and Paris (Data Reuse in the Humanities) are designed as experiments in developing a format which will engage advanced users and foster a collective reflection on "data fluidity" (see Romary, Mertens & Baillot 2016).

DARIAH & affiliates (aka "DARIAH's outer network")

Matej Durco

Despite the vast diversity of the individual research agendas, at a certain level all research infrastructures and initiatives need the same kind of basic facilities represented by the lower levels of the architectural stacks – stable hosting and storage of data, processing and computational capacity, secure authentication to build trust, stable referencing of resources, etc. DARIAH has always embraced the idea of cooperation between the different initiatives to harness the synergies emerging from this shared needs and has strong ties both "horizontally" to affiliated RI-projects (EHRI, CENDARI, PARTHENOS) and to sibling infrastructures like CLARIN, as well as "vertically" to networks that specialize on the actual provision of these basic services like eduGain, EGI, or EUDAT. This is in line with DARIAH's self-image as a facilitator that speaks the language of the researchers, but also has the "global critical mass" and expertise to approach and negotiate with large technical infrastructures.

Bibliography

Borgman, C.L., Edwards, P.N., Jackson, S.J., Chalmers, M.K., Bowker, G.C., Ribes, D., Burton, M. (2013) *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Report. http://pne.people.si.umich.edu/PDF/Edwards_etal_2013_Knowledge_Infrastructures.pdf

Doorn, P., Aerts, P., Lusher, S. (2016) *Research Software at the Heart of Discovery*. Technical Report, DANS, NLeSC, 2016. https://www.esciencecenter.nl/pdf/Software_Sustainability_DANS_NLeSC_2016.pdf

Fihn, J., Gnadt, T., Hoogerwerf, M., Jerlehag, B., Lenkiewicz, P., Priddy, M., Shepherdson, J., *DASISH - Reference Model for Social Science and Humanities Data Infrastructures*. European Commission, 2016. https://sites.google.com/a/dans.knaw.nl/reference-model-for-ssh-data-infrastructure/

Romary, L., Mertens, M. et Baillot, A. (2016) *Data Fluidity in DARIAH – Pushing the Agenda Forward*, *Bibliothek Forschung und Praxis*, 39, 3, 2016, p. 350–357, https://hal.inria.fr/hal-01285917.

Wouters, P., Beaulieu, A., Scharnhorst, A., & Wyatt, S. (eds) (2013). *Virtual knowledge: Experimenting in the humanities and the social sciences*. Cambridge, MA: The MIT Press.

Only Connect!: Intertextuality, Circulation, and Networks in Digital Resources for Women's Writing

Alison Booth
ab6j@eservices.virginia.edu
University of Virginia, United States of America

Sarah Connell
sa.connell@neu.edu
Northeastern University, United States of America

Marie-Louise Coolahan
marielouise.coolahan@nuigalway.ie
National University of Ireland, Galway, Ireland

Jeremy Boggs
jeremey@virginia.edu
University of Virginia, United States of America

Julia Flanders
j.flanders@neu.edu
Northeastern University, United States of America

David Kelly
david.d.kelly@nuigalway.ie
National University of Ireland, Galway, Ireland

Rennie Mapp
rcm7e@virginia.edu
University of Virginia, United States of America

Worthy Martin
martin@virginia.edu
University of Virginia, United States of America

Introduction

This panel brings together papers by three projects that are working on digital methods for researching and representing the textual networks of women writers. "Documentary Social Networks: Women Travel Writers in Prosopographies," by Alison Booth with Worthy Martin, Rennie Mapp, and Jeremy Boggs, focuses on a cohort of travel writers from the Collective Biographies of Women project to consider documentary social networks and the "sibling" relationship of subjects who occupy a single biographical volume. "Intertextual Networks: Theorizing and Encoding Textual Connections in Early Women's Writing," by Sarah Connell and Julia Flanders, presents the Women Writers Project's (WWP) work on using markup to model and research the citation and quotation practices of early women writers in the context of the larger ecology of digital projects focused on gender. "Digital Representation, Intertextual Relationships, and the Impact of Early Modern Women's Writing," by Marie-Louise Coolahan and David Kelly, describes the research of the RECIRC project (The Reception and Circulation of Early Modern Women's Writing: 1550–1700), focusing on RECIRC's development of taxonomies of reception and circulation and on the methodologies for understanding textual transmission that the project has been testing.

These projects explore the problem of intertextuality at several different levels: primary source documents (including print and manuscript), metadata, and synthetic or critical narratives such as biography and annotation. Each project faces the challenges particular to its own place in the digital ecology and also the challenge of making its data usable in other parts of that ecology: connections that will depend on the collaborative generation of digital standards. The projects also share an interest in understanding the roles that gender played in the circulation of texts, investigating what women read and cited, which women were read and cited, and how women were represented in writing. Together they offer an opportunity to explore how intertextuality, as understood in the context of the modern web of linked open data, operates not only as a deliberate rhetoric of quotation but also through more complex interconnections between texts, authors, and other cultural agencies—and how those links can effectively be brokered through connections between digital research projects and resources.

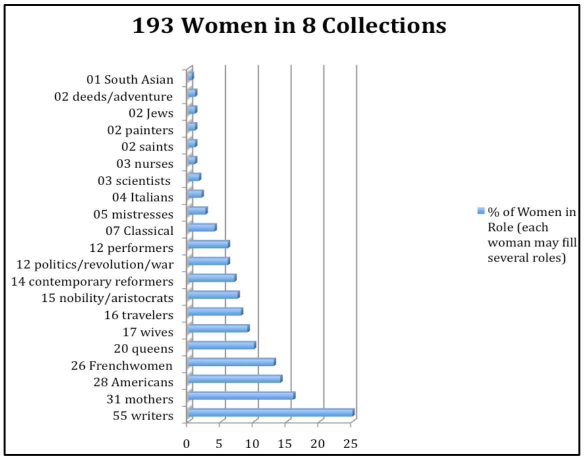
These three papers will not only discuss the various textual networks of women writers that are the subject of their individual projects but will also consider connections between the projects themselves, highlighting both points of contact and distinctions. We see this panel as an opportunity to examine the wider possibilities for making intertextual connections legible among digital systems of textual circulation.

Documentary Social Networks: Women Travel Writers in Prosopographies

Alison Booth with Worthy Martin, Rennie Mapp, and Jeremy Boggs

New literary histories using large digitized corpora can scale up the longstanding interest in intertextuality, beyond allusions and acknowledged influences. Such projects as the Women Writers Project and RECIRC engage not only textual but also personographic data, as all kinds of metadata interdependently form rigorous measures of intertextual transmission. As feminist literary history, such projects are also committed to contexts of women's lives and networks of representation. The Collective Biographies of Women (CBW) project studies Anglophone prosopographies (collective biographies, or a quantitative approach to standardized lives in sets) women of all occupations in books published since 1830, revealing trends in representation of women through an annotated bibliography, a database and stand-aside XML markup of narrative structure and discourse (Biographical Elements and Structure Schema, or BESS a stand-aside XML schema for marking elements of biographical narratives at the level of the paragraph). Unlike many feminist DH projects, CBW is not an archive of women's creative writing, though at least 700 of 8700 persons in the database are writers (one person is likely to represent several types). We study the collocation of women in books; thus Harriet Martineau, in 25 collections, is a "sibling" in "one degree of separation" from the 316 individuals found at least once in those collections. We have begun with typological cohorts and sample corpora focused on a nodal woman (all the volumes in which her biography appears) because of the labor-intensive analysis in our project's design.

This talk introduces the cohort of travel writers in CBW, focusing on the sample corpus of volumes that include a short biography of the famous mother and travel writer Frances Trollope (CBW includes 10 books entirely of female travelers; 17 collections of great mothers or mothers of the great. It should be noted that CBW assigns 137 collection types– e.g. genre; theme; types of women in them– and the 1272 volumes may each be assigned more than one type). Her biography joins assortments of different types:



One exemplary text for this paper will be Una Pope-Hennessy's Three Englishwomen in America (London: Benn, 1929), on Frances Trollope (for an earlier account of Frances Trollope as travel writer and mother, see Booth, 2015), Harriet Martineau, and Fanny Kemble, to illustrate the concept of documentary social networks and the "siblings" who occupy a single volume. More recently, we study the author of such a collection, in the cohort of some 995 "presenters" (i.e. biographers, illustrators, editors), 567 of whom were female. Currently, the CBW database relates an author such as Pope-Hennessy only to the book, not to the persons associated with its chapters. We would like to visualize the relations among authors and their subjects as well as we already trace intertextual exchange among versions of the same lives.

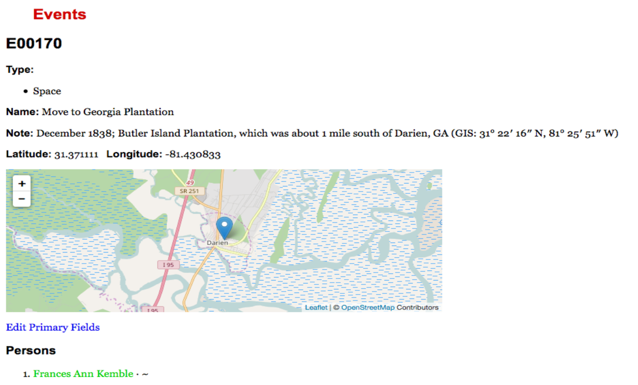
DAME UNA POPE-HENNESSY
THREE
ENGLISH WOMEN IN
AMERICA
Fanny Trollope Fanny Kemble Harriet Martineau

Kemble, Trollope, Martineau, circa 1830s (Jameson, 1844)

Figures 1,2. Kemble: Steel engraving by Johnson Wilson, & Co., after painting by Alonzo Chappel, after painting by Sir Thomas Lawrence. Trollope: Frontispiece: "Francis Trollope from a portrait painted by A. Hervieu," in Frances Eleanor Trollope, Frances Trollope: Her Life and Literary Work from George III. to Victoria, 2 vols. (London: Bentley, 1895). Harriet Martineau by Richard Evans (National Portrait Gallery). Mrs. Anna Brownell Jameson, 1844, Hill and Adamson salt print, Art Institute of Chicago.

The portraits of British women authors who wrote famous observations on America in the 1830s, adapted as frontispieces, take on a family resemblance in spite of the great social and personal distance among these mutual acquaintances. Covers and illustrations often create composite portraits (see Booth, 2004:33–42) Pope-Hennessy's Englishwomen had bestselling books or tours in the 1830s, warranting their then-fashionable portraits (Cecil Beaton photographed Pope-Hennessy). Whereas Trollope and Kemble, and Martineau and Kemble, only coincide in Three Englishwomen in America, Martineau and Trollope share 8 collections, 3 on travelers, others on more miscellaneous achievement. What common ground justifies narratives of their lives in a single publication? Pope-Hennessy's title suggests it is their gendered transatlantic vision. It does not spell out that these Victorian writers attacked America for its manners and for slavery. CBW finds distinctive features in lives of women writers who traveled, besides an increase in geospatial data: a publication event characterizes and types the author ("Trollopize" is the verb the came from Trollope's offensive Domestic Manners of the Americans); literary biographies devote a high proportion of paragraphs to summary of the works, in contrast with biographies of queens or nurses, for example.

Databases tend to label persons by nationality and familial role. Our research shows the temporal or narrative dimension of supposedly fixed traits as well as the unstable scope of events. Englishness (the event "birth") should be modified by later adopted countries; Trollope and Kemble emigrated to the US for some years, and both primarily resided in Italy in later life. Married mothers, novelist Trollope and actress Kemble later separated or divorced; political theorist Martineau was unmarried and childless; these relations shape their representation. Many events challenge the researcher to affix GIS or standard dates, such as the bankruptcies (tied to family relations) that enabled all three writing careers; these crises had far-reaching continuance in creating their opportunities or need to travel. We have identified the most frequently narrated events in all versions of the three writers' lives, potentially mapping their travels, works, and lives in a way adaptable for any intertextual biographical cohort.



Events

E00170

Type:

- Space

Name: Move to Georgia Plantation

Note: December 1838; Butler Island Plantation, which was about 1 mile south of Darien, GA (GIS: 31° 22′ 16″ N, 81° 25′ 51″ W)

Latitude: 31.371111 Longitude: -81.430833

Edit Primary Fields

Persons

1. Frances Ann Kemble · ~

The talk shares portraits, visualizations of documentary social networks, maps, and innovative interface (rich-prospect browsing) to reveal analyses of versions of the same person or persons within networks of text. We sample intertextuality in which these and other women enter into the body of each others' life narratives. Like any biography, Pope-Hennessy's chapters assemble citations and redactions of previous versions, from letters, archives, and autobiographical records to biographies; the textual methods of RECIRC or a tool like Juxta, as well as BESS analysis, fruitfully compare versions. The paper will reflect on the intertextuality of representation of women writers and travelers, as well as the challenges of a comprehensive prosopographical study using digital tools to develop a large-scale and finely grained analysis of women's biographical histories.

Acknowledgements

The Collective Biographies of Women Project is supported by the E-Text Center, Scholars' Lab, and Institute for Advanced Technology in the Humanities (IATH) as well as English Department at the University of Virginia, CBW has received ACLS Fellowship and NEH ODH Level-II Startup funding

Intertextual Networks: Theorizing and Encoding Textual Connections in Early Women's Writing

Sarah Connell and Julia Flanders

Introduction

Intertextual Networks is an initiative of the Women Writers Project (WWP) at Northeastern University aimed at exploring and theorizing the representation of intertextuality, with a focus on the citation and quotation practices of the authors represented in the WWP's digital collection, Women Writers Online (WWO). The WWP's work on Intertextual Networks incorporates several strands: focused projects conducted by individual research collaborators; sustained examination of the modalities of intertextuality as revealed by the work of our staff and collaborators; and a large-scale encoding project creating a bibliography of all the texts named or quoted in WWO, linking the texts in that bibliography with their occurrences in the WWO corpus, and substantially expanding the encoding of intertextual phenomena in our textbase. In this paper, we will discuss the aims and methods of the project, offering models for encoding complex intertextual features and setting out some processes for the systematic application of additional markup to an existing corpus. We will also consider the implications of this project for the larger ecology of digitized collections focused on gender and on women's writing.

Contexts

While Intertextual Networks is a recent initiative of the WWP, it has grown out of several decades of previous research as manifested in WWO and other WWP publications–particularly *Women Writers in Review*, a collection of around 700 reviews and publication notices responding to the authors in WWO. The almost 400 texts published in Women Writers Online are primarily print English-language works, representing a broad cross-section of texts written and translated by women from 1526 to 1850. These texts have been transcribed and encoded using the Text Encoding Initiative (TEI) Guidelines. Intertextual Networks also builds on and contributes to ongoing research into the digital representation of intertextuality, including the substantial work already evident in the TEI Guidelines' recommendations for encoding titles, quotations, and other textual references. Additionally, we are working with the Orlando Project's (see also, Brown et al, 2004) and the RECIRC project's (The Reception and Circulation of Early Modern Women's Writing, 1550–1700) bodies of research into developing taxonomies of reception, circulation, and intertextuality. Intertextual Networks is equally grounded in literary and historical scholarship on the ways that women from the early modern period to the mid-nineteenth century read and responded to texts (e.g., Horrocks, 2008; Rumbold, 2006; Winterer, 2008).

Design and goals

This project is working to create a much clearer and more textured picture of the rhetoric of intertextuality:

what female authors read; what they felt it important to quote, paraphrase, or cite; and what mechanisms connect their writing to that of other authors. In addition to the relatively straightforward instances of explicit quotations, citations, and references to specific titles, the project is also invested in developing practices for marking up subtler forms of intertextual engagement that emerge from verbal echoes, stylistic or topic similarities, imitation, parody, and other transformative ways of responding to what one has read.

Because WWO is chronologically broad and generically diverse, it provides a considerable range of opportunities to test different encoding practices against textual exemplars. The textual references in WWO are often densely layered and quite complex—for just a few examples, Lady Eleanor Davies inserts the full text of other documents into some of her political pamphlets; the 1706 Ladies' Diary constructs short poems, called "Enigmas," out of lines from several other poetic works; and Elizabeth Craven's 1789 A Journey through the Crimea to Constantinople inserts a poem that she wrote based on, and sometimes "literally translated" from a pamphlet, with citations to the pamphlet itself (43). Thus, an important goal of this project is not only to reveal early women's intertextual practices but also to test and model methods for representing the considerable complexity of these practices in TEI markup, working with formal categories without flattening out useful levels of nuance.

Encoding and bibliographic development

In developing a bibliography for the texts named in WWO, we have found that a balance of programmatic intervention and human attention can effectively accomplish systematic adjustments across our corpus. Using XQuery, we have generated a spreadsheet with the distinct titles—and authorship details where they are available—referenced in WWO. The WWP's encoding staff has been gathering basic publication details (standardized titles, authors, and dates and locations of initial publication), removing duplicates, disambiguating wherever necessary—such as with the many texts that are titled "Poems"—and adding unique identifiers. Our bibliographic data follow Functional Requirements for Bibliographic Records (FRBR) recommendations. Preserving the titles as they are represented in WWO, with the XPaths used to locate each, means that we can automatically add the unique identifiers back into the corpus when we have completed the initial bibliographic work. This approach enables us to identify the more than 6,000 <title> elements in WWO relatively quickly, while ensuring from the outset that there are no duplicate entries and laying the groundwork for the future additions of texts and textual details as we expand our source identification to quotations and other textual references.

We have also found that human and programmatic intervention can be fruitfully combined in establishing and implementing methods for encoding those intertextual features that are not straightforward <quote>s and <title>s and in expanding our understanding of how quotations and titles are being used by the authors in WWO. XPath and XQuery can reveal usage patterns and identify cases that might warrant additional investigation—such as the relatively rare instances in which titles are named in dramatic verse. Corpus-wide queries are also useful in locating additional examples of textual phenomena in order to decide how best to encode them. For example, we have reviewed emendations and translations of quotations, indirect references to titles, instances in which textual materials are elided from quotations, quotations within quotations, and incorrect or partially correct citations to develop consistent encoding practices that are applicable across our corpus, despite its variety.

Conclusion

This project is concerned with both discovering textual reverberation (the traces of women's reading that emerge in their writing) and with making that reverberation legible within the new digital systems of textual circulation. Within the boundaries of the Women Writers Project, that legibility is effected through the encoding that makes intertextuality an explicit feature of our modeling of texts. The use of a community standard like the TEI and the future availability of an API to the project's data extend that legibility—in principle—beyond the project's walls, but these methods do not in themselves build the complex web of interconnections that would constitute digital intertextuality. By placing the WWP's work alongside that of Collective Biographies of Women and RECIRC (two out of a much wider field of relevant connections) this panel will suggest what that larger intertextuality could look like, and what further work would be needed to realize it.

Digital Representation, Intertextual Relationships, and the Impact of Early Modern Women's Writing

Marie-Louise Coolahan and David Kelly

This paper emerges from the research of the team working on the RECIRC project (The Reception and Circulation of Early Modern Women's Writing, 1550–1700), funded by the European Research Council (2014–2019) and led by Marie-Louise Coolahan. It will describe the project, its development of digital tools, collaborations, and plans for interoperability with cognate projects. It will focus in particular on our development of taxonomies of reception and circulation, designed to capture data that reflects early modern source material, as the basis for our dialogue with the Women Writers Project and Collective Biographies of Women project – a dialogue that is centrally concerned with questions of intertextuality, circulation and the collaborative generation of digital standards.

The RECIRC project is essentially a study of intellectual impact. Its fundamental research questions include: Which women were read? How, where, and by whom were they read? RECIRC is structured around four interlinking 'work packages', each of which takes a specific entry point in order to amass quantitative data relating to the reception and circulation of women's writing between 1550 and 1700. The first of these posits the Catholic religious orders as transnational channels by which devotional and polemical texts were translated and transmitted; it investigates the martyrologies and bibliographies of the various religious orders, as large-scale compendia of texts that included female-authored works. The second 'work package' examines scientific correspondence networks (and therefore also complements the research currently brought under the umbrella of Women's/Early Modern Letters Online (EMLO, WEMLO) and Reassembling the Republic of Letters (led by Howard Hotson); the wealth of data to be found in the scriptorium operated through Samuel Hartlib has meant we have focused specifically on this circle. The third approach aims to rebalance the bias of digitization projects toward print culture by harvesting data from early modern manuscripts. It does so by focusing solely on the category of the manuscript miscellany (a compilation of miscellaneous materials) in order to assess the contexts for excerpting and transcribing women's writing. It differs from the Folger Shakespeare Library's Early Modern Manuscripts Online (EMMO) initiative, which is a full-text transcription project, in its harvesting and structuring of data relating specifically to reception and circulation. The fourth RECIRC approach is concerned with early modern library catalogues; it captures data on the proportion of female-authored items in order to facilitate statistical analysis relating to the gendering of such book collections.

RECIRC, then, is testing these methodological approaches for understanding the 'big picture' of textual transmission, reception and circulation of women's writing in the sixteenth and seventeenth centuries. The focus on women's writing enables investigation of the routes to impact that were exploited by early modern women, as well as of the ways gender inflected the construction of writerly reputation. It also delimits the corpus, facilitating our testing of methodologies for studying the circulation of non-elite, non-canonical writing in the period. Rather than producing a full-text digitization of primary materials, the project has been centrally concerned with developing taxonomies of reception and circulation – and these are the basis of collaboration with the WWP's current Intertextual Networks project and the Collective Biographies of Women project. This encompasses productive conversations around the kinds of intertextuality (quotation, excepting, citation of text and/or author) that occur in relation to women's writing, and their modes of digital representation. There are also important areas of divergence: RECIRC is working with metadata categories rather than xml tags; although each instance of reception evidence is supplied as full-text, these instances are themselves extracts from significantly larger texts. Moreover, RECIRC is concerned with all – women's and men's – reception of female authors, which allows for equally productive conversations about gender and reception.

RECIRC data are stored in an online database, which will be made publicly accessible at the project's close, and is intended to be interoperable with cognate projects, such as the NEWW Women Writers Virtual Research Environment. The database architecture (built using a RESTful API approach) enables multiple output formats and we will discuss possibilities for interoperable outputs. Moreover, the project is now (October 2016) at the stage of data cleaning, in preparation for experimenting with visualization tools and quantitative analysis. We aim to create network visualizations and analyses that embrace both the gendering of reception and the relationships of texts with each other. Questions include: Which genres of female-authored texts were most popularly circulated? What forms of circulation (translation, excerpts, citation) were most conducive to their transmission? How important (and prevalent) was attribution to their circulation? Which female authors were reading and using other women writers? What circulation contexts promoted women as authors? If accepted to DH2017, we intend to present our preliminary answers to the questions and patterns that emerge during this quantitative analysis phase.

Bibliography

Booth, A. (2004). How to Make It as a Woman: Collective Biographical History from Victoria to the Present. Chicago: U Chicago Press.

Booth, A. (2015). "Frances Trollope in a Victorian Network of Women's Biographies." Virtual Victorians: Networks, Connections, Technologies, ed. Veronica Alfano and Andrew Stauffer. Palgrave Macmillan.83–106. Digital Annex http://www.virtualvictorians.org/

Brown, S., Grundy, I., Clements, P., Elio, R., Balzas, S., and Cameron, R. (2004). Intertextual Encoding in the Writing of Women's Literary History. Computers and the Humanities. 38(2): 191–206.

Craven, E. (1789) A Journey through the Crimea to Constantinople. London: G. G. J. and J. Robinson.

Early Modern Letters Online (n.d.) http://emlo.bodleian.ox.ac.uk

Early Modern Manuscripts Online (n.d) : http://folgerpedia.folger.edu/Early_Modern_Manuscripts_Online.

Horrocks. E. (2008) 'Her Ideas Arranged Themselves': Re-Membering Poetry in Radcliffe. Studies in Romanticism 47(4): 505–527.

IFLA Study Group on the Functional Requirements of Bibliographic Records. (1998). Functional Requirements of Bibliographic Records: final report. Last updated February 2009. http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.ht

Martineau, H. (2007). Autobiography, ed. Linda H. Peterson (Peterborough, Ontario, Canada: Broadview).

New Approaches to European Women Writers (NEWW) (n.d.) Women Writers VRE: http://resources.huygens.knaw.nl/womenwriters

Reassembling the Republic of Letters: http://www.republicofletters.net

Pope-Hennessy, U. (1929). Three Englishwomen in America (London: Benn,). http://cbw.iath.virginia.edu/books_display.php?id=1990.

Rumbold, K. (2006). .'Alas, poor Yorick': Quoting Shakespeare in the Mid-Eighteenth-Century Novel. Borrowers and Lenders: The Journal of Shakespeare and Appropriation 2(2). http://www.borrowers.uga.edu/7151/toc

TEI Consortium. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.0.0. Last updated 29th March 2016. http://tei-c.org/release/doc/tei-p5-doc/en/html/

Winterer, C. (2008) The Female World of Classical Reading in Eighteenth-Century America. In Brayman Hackel, H. and Kelly, C. (eds.), Reading Women: Literacy, Authorship, and Culture in the Atlantic World, 1500–1800. Philadelphia: University of Pennsylvania Press. 105–123.

Hack, Yack, Stack: Access, Culture, and DH Infrastructure

Laura Braunstein
lrb@dartmouth.edu
Dartmouth College, United States of America

Thomas Padilla
thomaspadilla@ucsb.edu
UC Santa Barbara, United States of America

Amanda Visconti
amandavisconti@gmail.com
Scholar's Lab, University of Virginia Library,
United States of America

Abstract

In an essay exploring the possibilities of "critical infrastructure studies," Alan Liu calls upon scholars to "'see through' the supposed rationality of organizations and their supporting infrastructures to the fact that they are indeed social institutions with all the irrationality that implies." This panel brings together three librarians from different institutional contexts to interrogate DH infrastructure as a primarily social -- and irrational -- formation. If DH infrastructure is "only" what we make of it, how do we consciously, actively shape DH infrastructure to support what DH could be? How do the social formations of DH infrastructure activate -- but also potentially inhibit -- access by scholars, students, and publics? How do certain social formations of institutional DH mask the emotional labor of DH professionals supporting the work of other researchers? And finally, what is the relationship between infrastructure and underlying institutional culture? Can a critically engaged infrastructure overcome or repair a culture that may undermine the values that we want to instill?

Laura Braunstein will explore the challenges of sustaining DH infrastructure without a DH center. Project management, collaboration, outreach, and digital pedagogy take an ad-hoc and at times, rogue presence, connected only by virtual infrastructure. A DH community of practice that exists primarily as a social network can enable serendipitous connections, but it is not sustainable, for scholars, teams, or projects. How can we institutionalize a virtual DH non-center in order to support activities and projects without inhibiting its energy and possibility?

Thomas Padilla will explore assumptions latent in prevailing notions of scalability in the Digital Humanities. The need to qualify DH infrastructure development, web-based projects, and data sources by the extent to which they are scalable or not tends have the effect of either totally disincentivizing participation by cultural heritage organizations or incentivizing in a manner that tends to occlude acts of individual agency and creation. How can we work to develop practices and methods for exposing and according value to a wide range of cultural heritage work? In what ways does exposing agency and according value to intellectual and physical labor a necessary predicate of research integrity in DH?

Amanda Visconti will explore how we might align DH resource investment with explicit personal and intellectual values such as inclusion and open access. Multiple DH centers already define the scope of the DH they support not through some over-generalization (e.g. "we support innovation in humanities via digital scholarship"), but instead through charters envisioning and encouraging a DH that performs what we care about. How do we use the wiggle room at our disposal, small or large, to push our institutions toward supporting a better DH? How can we act on our values through policies that center resources on work strengthening these values?

Panel organization

As the discussion is informed by each speaker's local context, each of the three speakers will give a 5-minute introduction of their background and interests (15 minutes total). Each speaker will then interview the other two speakers with a prepared question (2 questions from each speaker, 6 questions total at 5 minutes each, 30 minutes total). At the last part of the panel, we'll open up the discussion for both Q&A and audience sharing; the panelists will moderate the conversation so that everyone gets a chance to contribute.

The speakers are:

- Laura Braunstein, Digital Humanities Librarian, Dartmouth College

- Thomas Padilla, Humanities Data Curator, University of California Santa Barbara
- Amanda Visconti, Managing Director, Scholars' Lab, University of Virginia Library

Bibliography

Braunstein, L., and Kim, J. (2016). "6 Questions for a Digital Humanities Librarian". Inside Higher Ed. August 17, 2016. https://www.insidehighered.com/blogs/technology-and-learning/6-questions-digital-humanities-librarian

Liu, A. (2016). "Drafts for Against the Cultural Singularity (book in progress)." May 2, 2016. http://liu.english.ucsb.edu/drafts-for-against-the-cultural-singularity

Mandell, L., and Dinsman, M. (2016) "The Digital in the Humanities: An Interview with Laura Mandell". L.A. Review of Books, April 24, 2016. https://lareviewofbooks.org/article/digital-humanities-interview-laura-mandell

Morgan, P. (2016). "Not Your DH Teddy Bear; or, Emotional Labor is Not Going Away.' DH+Lib, July 29, 2016. http://acrl.ala.org/dh/2016/07/29/not-your-dh-teddy-bear/

Nowviskie, B., and Dinsman, M. (2016). "The Digital in the Humanities: An Interview with Bethany Nowviskie". L.A. Review of Books, May 9, 2016. https://lareviewofbooks.org/article/digital-humanities-interview-bethany-nowviskie

Padilla, T., and Peet, L. (2016) "Thomas Padilla, UCSB's Inaugural Humanities Data Curator". September 15, 2016. http://lj.libraryjournal.com/2016/09/people/thomas-padilla-ucsbs-inaugural-humanities-data-curator

Padilla, T. (2016). "Collections as Data: Conditions of Possibility". October 28, 2016. https://medium.com/@tgpadillajr/collections-as-data-conditions-of-possibility-494805bf16be#.8bvuflwft

Sayers, J. (2016) (@jenterysayers) "Makerspaces are culture first, infrastructure second. The culture is often toxic + exclusive and warrants more social justice research." October 16, 2016, 5:18pm. Tweet.

Visconti, A. (2016). "Designing a Digital Humanities Initiative: Background & Campus DH Survey". Literature Geek blog, September 9, 2016. http://literaturegeek.com/2016/09/11/designing-digital-humanities-initiative-background-campus-survey

Treating a Genre as a Database: the Chinese Local Gazetteers, the LG Tools, and Research Based on This New Digital Methodology

Shih-Pei Chen
schen@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Qun Che
qche@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Ling Cao
caoling1013@163.com
Nanjing University of Information Science & Technology China

Dagmar Schäfer
dschaefer@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Hongsu Wang
hongsuwang@fas.harvard.edu
Harvard University, United States of America

This panel discusses how digitization and digital tools help to bring new insights to a well studied genre, in this case the Chinese Local Gazetteers, by supporting research inquiries that treat the whole genre as a conceptual "database" to answer especially large scale questions that take into account gazetteers from multiple geographic regions and within long time spans.

The Chinese Local Gazetteers is a long established genre of writing in China since the twelve century for recording local knowledge about a region. Local gentry and officials compiled information about a region, ranging from landscape, flora and fauna, officials and celebrities to temples and schools, local culture and customs, and taxes and census, and kept them in this genre. Despite that Local Gazetteers have been major sources for scholars to find specific information about a place, it turns out to be very difficult if a scholar wishes to study the gazetteers on larger scales due to the vast amount of information contained within. Thanks to the increasing recognition of digitizing historical sources, as of 2016 nearly half of the extant 8,000 titles of local gazetteers (one title can mean one to dozens of physical volumes) has been digitized as searchable full texts and provided access through various databases from public and private sectors. However, despite of the large amount of local gazetteers available electronically, these databases mainly provide features that replicate how physical books are used: users are able to flip a gazetteer page by page; even when reading the returns from a full text search, a scholar still needs to click the returns one by one in order to read the texts. Such principle of treating the gazetteers individually highly restricts the possibility of researching large sets of digital gazetteers. Historians need better tools to work with the large amount of digitized gazetteers to employ new forms of digital research methodologies and to fully benefit from the digital formats. One important methodology that we address here is the possibility to treat the whole genre as the body of inquiry in order to research global and large scale phenomena that are across individual gazetteers, geographical regions, and time spans.

Under this context, Department III “Artefacts, Actions, and Knowledge” of the Max Planck Institute for the History of Science (MPIWG) embarks on a digital project to build digital tools specifically for this genre that allow scholars to adopt this methodology of treating the available set of digitized gazetteers as a conceptual “database” to post inquiries. The digital tools, which we call **the LG Tools**, include a full text search facility, an Extraction Interface to collect data in the form of lists, a research repository to store and publish collected data, and an interactive mapping and analysis platform that are linked to data collected via full text search or the Extraction Interface in order to visualize the data geographically. We have presented the LG Tools last year in this conference. The abstract, which contains detailed description for the tools, can be found online.

In this panel, we would like to shift the focus from the tools themselves to the types of research that can be derived from the LG Tools and from the methodology of treating the whole genre as a database. We invite four scholars in Chinese history to demonstrate their research projects and to discuss what advancement the LG Tools have brought to their research. The topics of their research range from history of science and technology, environmental history, to social and intellectual histories. The four historians and their research projects are described below.

Qun Che

Qun Che uses the LG Tools to trace the transformation of the geographic landform of the Dongting Lake, a flood basin of the Yangtze River, by looking at the construction of “Yuan” recorded in the local gazetteers during the Ming and Qing dynasties (1368-1911). Yuan is a type of water conservancy that literally refers to fields surrounded by embankments to prevent river flooding. Che first finds all the occurrences of Yuan constructions in the available digital set of local gazetteers by full text search. The resulting dataset is then sent to LGMap, the interactive mapping and analysis interface among LG Tools. By using LGMap, Che is able to clarify the Yuan constructions in the Dongting Lake region in different time periods and by doing so to trace the flooding and deposition process of the lake to understand how the lake has been shaped from the past to today. Figure 1 shows the LGMap interface for analyzing the periods and locations of the Yuan constructions near the lake.

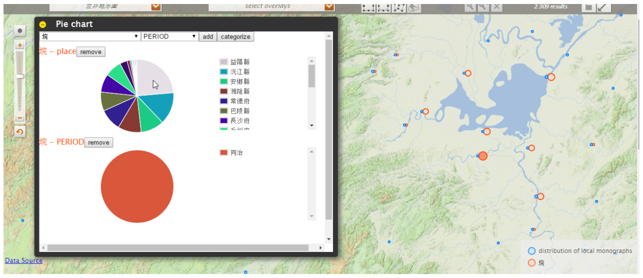


Figure 1. Yuan constructions recorded in the available digital set of Local Gazetteers, visualized in LGMap

Ling Cao

Ling Cao collects records in the local gazetteers describing the growing and spreading of maize in China during the Ming and Qing dynasties in order to understand through which geographical routes maize was introduced to and spread in China. She further relates the regions of maize growing and their scales to the occurrence rate of natural disasters of the typical maize regions to see the influence of maize over the ecological environment. From the records she collects, Cao is able to raise the hypothesis that the introduction and growing of maize in the mountain areas might cause the destruction of forests and result in serious soil erosion which in turn led to blockage in downstream rivers and floods.

Dagmar Schäfer

Dagmar Schäfer researches on the relationship between material availability and the development of local knowledge organization in the local gazetteers. While not all local gazetteers share the same classification schemes or even topic, local officials recorded local expertise and material specialties in the gazetteers. In which sections did local gazetteer compilers place such information? When and how did their descriptions vary (such as details, terminology, etc.)? The full text function of the LG Tools and the section titles that MPIWG invested to type up give hints on the above questions. The contribution also aims at discussing the possibility of tracing and analyzing networks of materials: how did officials relate materials to another; can scholars identify clusters among the materials. The contribution also posts questions on how social network analysis can be applied and what are the advantages of geographical mapping.

Hongsu Wang

Hongsu Wang will demonstrate how the LG Tools especially the Extraction Interface helps the China Biographical Database project (CBDB) to collect data about local officials from local gazetteers. CBDB is a freely accessible relational database with biographical information about more than 360,000 individuals in historical China. In addition to the large dataset, CBDB embeds functions to analyze the individuals via statistical, social network, and spatial analysis. CBDB turned to local gazetteers when it tried to increase its collection on individuals that were mainly based on central government records. The individuals listed in the sections of "local officials" in most gazetteers are ideal for this purpose since most of them don't appear in imperial government records. By using the Extraction Interface, which provides a regular expressions editor to help capture regularly written information such as lists, student assistants at CBDB were able to collect 250,000 officials from 291 local gazetteers using just 420 man-hours.

Organization of the panel

This panel will include five speakers. To give the audience a background for discussion, Shih-Pei Chen (MPIWG) will begin by giving an introduction to the genre of Chinese Local Gazetteers and its characteristics as well as the LG Tools (15 minutes). Then, the other four panelists will each present their research projects and show how they get new findings through using the LG Tools (each speaker 10-12 minutes). We will then have a discussion session (roughly 30 minutes) for the panelists to talk about how the existence of the LG Tools has changed the objects of study and/or the approach to these objects and to discuss the conceptual frameworks and assumptions challenged by the use of digitized sources. We will also address the point of how the LG Tools help the panelists to develop the idea of a regarding the genre as a "conceptual database". The audience is invited to join in the discussion session for conversations on the types of advancement these tools bring to the study of history and humanities, whether similar tools can be set up for digital resources in other humanities disciplines, and the copyright issues encountered for building such tools.

Challenges for New Infrastructures and Paradigms in DH Curricular Program Development

Tanya Clement
tclement@ischool.utexas.edu
University of Texas at Austin, United States of America

Alison Booth
booth@virginia.edu
University of Virginia, United States of America

Ryan Cordell
r.cordell@northeastern.edu
Northeastern University, United States of America

Miriam Posner
mposner@humnet.ucla.edu
UC Los Angeles, United States of America

Maria Sachiko Cecire
mcecire@bard.edu
Bard College, United States of America

Introduction

Many leading universities in the United States have recognized the profoundly transformative effect that DH has had on research and teaching and have established lab-based research programs or institutional centers for interdisciplinary collaboration on digital projects in the humanities. The University of Virginia led with the Institute for Advanced Technology in the Humanities, which now supports more than forty DH research projects. The Stanford Humanities Laboratory, established in 2001, is a collaborative research environment for cross-disciplinary and multi-institutional, technologically transformative projects. Duke, a founding member of HASTAC (Humanities, Arts, Science, and Technology Advanced Collaboratory), adopted a similar model within its John Hope Franklin Humanities Institute. Other universities such as Harvard, Brown, Dartmouth, Berkeley, Princeton, and the University of Michigan have begun aggressively to hire in the field and to design multidisciplinary DH programs. Moreover, University of Pennsylvania and Yale have attracted large donations from alumnae for DH institutes, while state institutions like Maryland, Nebraska, and UCLA have garnered millions of dollars in external funds to support their field-leading digital scholarship centers.

Digital Humanities scholarship is, by necessity, collaborative and interdisciplinary. DH approaches to the creation and dissemination of scholarship investigate the information life cycle from digitization to preservation including multimediality, design elements, and computational reasoning and implementation through perspectives developed in the context of deep understandings of the key issues at stake in the humanities. DH approaches to teaching typically involve an emphasis on experiential learning and research by creatively expanding modes of access and networks of participation through the methods that students employ. This multidisciplinary work in research and teaching is not limited to conventional humanities departments, but, rather, emerges in every humanistic field, in arts and architecture, information studies, film and media studies, archaeology, geography, ethnic studies, and the social sciences. Because of this essentially multidisciplinary nature of DH work, some of the universities listed above have begun to move away from more traditional models of DH where DH is housed in the library or where DH is primarily an endeavor of the English or History Department. Northeastern University's center for Digital Humanities and Computational Social Science (NULab) and King's College DH Department, for example, bridge the humanities and computational, qualitative social sciences while Bard College's Experimental Humanities Concentration and Initiative is focused primarily on the arts.

Even with all this funding and enthusiasm, however, it remains surprisingly challenging to design curricula in DH on the undergraduate and graduate level. These challenges are a direct result of the fact that attempts to develop new paradigms and models for "training" students break (or stretch) molds for learning and teaching that have calcified over decades into departmental silos and administrative policies concerning credits and teaching loads. These challenges include the difficulty of organizing teaching arrangements that include faculty from multiple disciplines, supporting cluster hires versus supporting "reskilling" faculty to teach new methods from different perspectives, and designing single courses with heavy loads of material including perspectives on the world (such as cultural critique) alongside advanced topics in, for example, statistical models, computational approaches, and data visualization.

This panel will include representatives from five programs who bring perspectives from both private and public, large research universities and small liberal arts colleges in the United States. The panelists have worked within the context of long-standing digital humanities centers as well as within initiatives just now developing without these historical, infrastructural legacies. Each panelist will speak for twelve to fifteen minutes on the below topics, leaving time for questions from and conversations with the audience.

Tanya Clement

For many years, Digital Humanities at the University of Texas at Austin has happened on a project-level basis, without the support of a DH center, in American Studies, Anthropology, Classics, English, French and Italian, History, Information Studies, and in Portuguese and Spanish, among other departments (some projects can be viewed at the UT DH page)

At UT, we are experiencing a scenario that is familiar at universities and colleges that are supported by instantiated DH centers and initiatives. The existing, funded projects at UT do not offer enough project-based, training opportunities for a wide range of students. Project-based work with digital information technologies is often heralded as the site of work that defines DH (Burdick, et. al, 2012; Drucker, 2012; Hayles, 2012; Svensson, 2009, 2012). It is also the site of experiential learning and project-based research in DH on the undergraduate and graduate level, a trend that is noticeable in many of the lessons and curricula collected in the recent Modern Language Association publication, *Digital Pedagogy in the Humanities: Concepts, Models, and Experiments.* Even with this fabulous array of DH classroom-based assignments, the students who are best trained in DH -- those who on the graduate level are best prepared for the job market in private and public industry as well as the academy -- remain those students who participate in the kinds of long-term, interdisciplinary-team-led projects that have become the mainstay of DH. These kinds of projects, however, usually only have the funding to train a handful of lucky students. At UT, we are working to consider a different paradigm for teaching and training students that is not dependent, primarily, on the soft-money world that remains the status quo for most underfunded DH projects.

To better understand our proposed methods of training, it is important to describe the UT landscape. The University of Texas at Austin is a large research university with over 3,000 teaching faculty and over 50,000 students. It is rich in multicultural, multimedia, obscure and popular special collections of interest to humanities scholars. Some examples include the Gloria Anzaldúa archives, the Guatemalan National Police Historical Archive (AHPN), and the Radio Venceremos collection of digital audio recordings of guerrilla radio from the civil war in El Salvador housed in the Benson Latin American Collection. The papers of Carson McCullers, David Foster Wallace, and Gabriel Garcia Marquez as well as many collections of medieval and early modern manuscripts including one of five copies of the Gutenberg Bible in the United States are at the Harry Ransom Center, which also holds a robust collection of film (including the archive of Robert Dinero), photographs (such as the Magnum Photos' New York bureau collection, dating from 1929 to 2004), and authors' recordings including tapes of Anne Sexton's therapy sessions and Spalding Gray's performances alongside the center's other robust collections of twentieth century writers and performing artists. At the Briscoe Center for American History there is the Texas Poster Art Collection which documents the visuals behind the pivotal early music careers (1960s to 1970s) of iconic music legends such as Willie Nelson and Townes Van Zandt as well as over 2000 reels of tape that include the field recordings of folklorists John A. Lomax, William A. Owens, Américo Paredes, and John Henry Faulk.

The list of amazing collections at UT goes on, but in terms of DH, it doesn't matter. Most of these materials, like most of the materials (no matter if they are paper, reels, tape, or photographs) in special collections at many institutions are almost completely inaccessible to DH methods of presentation and inquiry – *even when they are digitized*. These materials are either under copyright; in non-text formats such as an image, audio, or video file; undescribed (and therefore, in many cases, undiscoverable); or unstructured; all of which make them unsuitable for most DH tools. Instead, UT scholars and teachers, like scholars and teachers everywhere, typically use the same collections that others use: what is freely available online – a trend that systematically limits the kinds of questions we can ask in DH scholarship and teaching.

At UT, we are trying to address these issues in our plans to create an undergraduate and graduate curriculum in DH by considering three primary questions that plague DH curriculum development everywhere:

1. How do we train a wide range of students on the undergraduate and graduate levels across a wide range of DH issues including, but not limited to, the creation of digital collections and archives; the analysis of digital materials; and the use of digital technologies to write, publish, and consume scholarship when our faculty are siloed from each other not only by administrative departments but by their, sometimes, many decades of differing experiences and training?
2. How do we teach students to consider multimedia and multiculturalism when using DH methods when the primary materials to which we have "data ready" access are text-based, in English, and unstructured?
3. How do we create an interdisciplinary framework that is productive and innovative as well as sustainable?

Building on the amazing work done by DH scholars in praxis-based training programs in libraries such as the Scholars Lab at the University of Virginia, Columbia's Developing Librarian Project, and Indiana University's Research Now: Cross Training for Digital Scholarship initiative, the DH@UT initiative is imagining a sustainable model for training students that is experiential, collaborative, and interdisciplinary. Collections are at the heart of humanities research. The work to make such collections "data ready" for DH scholarship coincides with deep expertise in the humanities as it relates to the organization, preservation, curation, analysis, visualization, and communication of digital works in the humanities. Pairing students with projects for making the collections at UT more accessible offers a unique opportunity to train students to generate a more diverse range of data-ready collections and to immerse them in questions surrounding critical information infrastructure studies (Clement, 2015; Liu 2016; Verhoeven 2016) that has engaged staff, undergraduates, graduates, and faculty at the heart of DH research.

Questions remain, however, about how such a program could be implemented. Faculty from a variety of disciplines will have to commit to using UT collections that perhaps do not fit exactly into their research objectives. The time commitments of library and archives staff, who are already often overwhelmed and under resourced, will need to be committed to the goals of the program; and, policies concerning how and when archival materials can cross the transom of the brick-and-mortar collection building (both physically and virtually due to copyright and privacy restrictions) will have to be reconsidered. There is reason to believe that faculty, used to "stretching" their area of expertise to teach classes, would be willing to commit their considerable effort in teaching classes in these new directions. There is equal reason to believe that staff members, committed to the goals of the institution to make their collections more accessible will also be in favor of such a program. Yet, in order to make this program sustainable, both faculty and staff will need support from upper administration. Commitment at the university level for innovative teaching and research remains imperative for changing the praxis of pedagogy. This talk will discuss our progress in these endeavors.

Alison Booth

Technological literacy is a stated educational goal for all students at the University of Virginia, and various existing groups in the Library and schools such as Arts and Sciences support teaching with technology and short assignments in courses (maps; e-portfolios, etc.; Learning Design Technology). But many would agree that digital humanities is something distinct from teaching with technology; there is more to be gained from student participation in open-ended projects (exceeding the timeframe of a single assignment) that reflect upon their tools and methods as well as on the specific data of a discipline. Considerable international discussion of DH pedagogy has advanced the field (Digital Pedagogy), and several textbooks serve courses that introduce students to DH (e.g. Gold and Klein). Certificates in DH (e.g. at Northeastern University and University of Maryland) provide models, as do departments of Digital Humanities (as at King's College, London). CenterNet presents a directory of DH centers on every continent, some like UCLA's Center for Digital Humanities offering a program with an undergraduate minor and graduate certificate.

What are the best models? How should we build an infrastructure for DH education at the University of Virginia, based on what we already have? The University of Virginia has longstanding centers practicing DH: the Institute for Advanced Technology in the Humanities, SHANTI, Scholars' Lab. Advanced research projects and courses or workshops in methods and tools are supported as well in Research Data Services, Data Sciences Institute, Makerspaces in arts, architecture, and engineering, among others. And yet the curricular offerings in DH have remained dispersed among a few academic departments and the Scholars' Lab's Praxis Program and fellowships, along with some digital fellowships for undergraduates working in art and archeology. I suggest a model: the Pedagogical Pyramid, which can be multiplied within one or many institutions in communities of interaction.

Is a Pedagogical Pyramid a menacing structure, an image of hierarchy and exploitative labor? Instead, it is intended as a metaphor and visualization (equilateral triangle in multiple dimensions) of a *graduated* structure of collaborations, potentially across institutions as well as UVA schools, on advanced research in the humanities, arts, and social sciences: more undergraduates, fewer graduate students, and fewer faculty. The numbers, limited only by practicable scales of collaboration, are flexible, but based on likely proportion of an institution's personnel in these ranks. The paid internships, fellowships, and mentorships proposed in this structure would be available in smaller numbers for team members who are at later career stages, but there will be no idle supervisory roles for those figuratively at the top. Participants would develop a charter in keeping with the Collaborators' Bill of Rights. With sufficient resources for faculty as well as students (stipends, wages, facilitators and spaces), we would run concurrent Pyramids.

We have developed parts of such a vision. IATH supports two residential fellows per year for two years each, with some course release and research funds. The Scholars' Lab under Bethany Nowviskie created a custom-built, annual cohort of six doctoral fellows who collaborate on a project. Praxis flourished under Purdom Lindblad and current members of Scholars' Lab, and has generated a Praxis Network connecting various institutions. Each year in addition there are 2-3 dissertation fellows in the Scholars' Lab, and we have had graduate fellows jointly in Data Sciences and Scholars' Lab. Undergraduates and graduates are employed in the Scholars' Lab Makerspace, and any students or faculty may work on any projects in that innovative research laboratory. We collaborate with a liberal arts college in Virginia, Washington & Lee, that adopted the Praxis model in its library-centered DH Studio; UVA and W&L hold exchanges among both institutions' Library staff, UVA's Praxis and DH fellows, and W&L's undergraduate students. Scholars' Lab has also led two summer programs for 4-6 Leadership Alliance Mellon Initiative (LAMI) undergraduate students from HCBUs and Puerto Rico, introducing DH and other research methods in preparation for graduate school applications. Collective Biographies of Women, an IATH and Scholars' Lab project, each year trains small groups of graduate research assistants (primarily MA) and a few paid undergraduates; LAMI students each summer have added research on African American and Latino cohorts.

Which brings us to planning the future of a DH community at UVA. Recent developments have focused on coordinating the community in a DH@UVA website and the DH@UVA 2016 conference, at which the open discussion came to a consensus on a certificate or program in DH. We do not want to build a DH department, as that may be less adaptable for future technological and curricular change.

Desirables: early introduction to digital research methods in interdisciplinary courses, workshops, and bootcamps suitable for undergraduates, graduate students, or both; engagement of undergraduates (for credit or pay) in faculty or graduate-student research projects in humanities, arts, and social sciences, beyond the classroom assignment; opportunities for graduate students to mentor and teach these undergraduates; paid graduate interns and project managers mentored by Library-affiliated faculty and staff who can work with faculty projects from many departments; different models of support for faculty digital scholarship beyond the IATH fellowships; expansion beyond the six Praxis students and 2-3 DH dissertation fellows per year that we currently support in Scholars' Lab.

Challenges for building such alliterative pyramids: persuading more faculty to participate and encourage their students in these opportunities; luring the CS faculty and students, the Data Sciences Institute, and the Library's Research Data Services group to collaborate with humanistic computational research, increasing broad interdisciplinarity; coordinating with the Graduate School and other schools to fund graduate fellowships, encourage dissertation advisors' participation, and monitor time-to-degree; securing additional resources for wages and faculty stipends as well as graduate fellowships and teaching release; negotiation with Directors of Undergraduate Studies and Graduate Studies as well as the Graduate School regarding requirements, credit courses and a certificate or program, which would have to be maneuvered through the channels of educational policy at university and state levels. Curriculum would have to be designed that allows discipline-specific units to be added to shared units on generally applicable tools, methods, and issues. The structures would have to ensure that all contributors are compensated and acknowledged, including in presentations and publications. UVA already has ample evidence of the enhancement of research and the advantages in students' learning and placement that come from the existing activities in DH groups. We look forward to giving these aims more substance and structure.

Ryan Cordell

During MLA 2013, Natalia Cecire wryly observed on Twitter, "1. DHers usually don't see dh as panacea. 2. Admins often do. 3. DHers often need for admins to have this erroneous belief." Our experiences building a graduate DH curriculum at Northeastern in many ways illustrate this rhetorical tension. We benefit from substantial administrative support for curricular ingenuity while struggling to reconcile that support with increasing disquiet in the departments that must underwrite any substantive changes we seek to make.

We are enormously fortunate at Northeastern and the NULab. Our administration has funded several years of cluster hires, which have allowed us to bring DH faculty into the English; History; and Cultures, Societies and Global Studies departments, as well as DH faculty and staff in the library. Over the course of four years we have founded a new center, integrated introductory and advanced DH courses into our curriculum, launched a DH graduate certificate program, and trained many students through work on locally- and grant-funded projects. This in turn has led to an increased number of students applying to our graduate programs seeking DH training.

This rapid, whole-cloth invention of a DH program, however, has been attended by pressures, fissures, and tensions with existing programs. For example, NULab faculty are proud of the robust coursework required for the DH certificate program: the equivalent of 3 courses out of the 10 required in our English MA program or 14 required in our English Ph.D., plus the development of a small scale DH project. Our students take not only an introductory DH course, but also advanced methods courses (data modeling, text analysis, etc.) that prepare them to integrate DH methods into their theses and compete for DH positions after graduation. Within English, however, completing this requirement requires students to decide their path almost immediately upon admission, and the decision to pursue the certificate dictates very particular paths through the larger Ph.D. program. While our DH faculty are a larger group than at most institutions, even so we cannot practically mount more than two courses per year: an introductory course each fall and an advanced course each spring. These advanced courses rotate among NULab faculty and thus have very distinct foci. Thus students' options for completing coursework remain relatively constrained over two years of full time coursework in ways that sometimes mitigate against the particular training individual's need. A student primarily interested in digital archive creation, for instance, might by necessity take their advanced course in Humanities Data Analysis rather than Data Modeling; while the latter would be more appropriate to their interests it can only be offered every three years or so, when a particular faculty member is on rotation for the advanced seminar.

These pressures are compounded for MA students in English or Public History; in the latter case we find there is really only one viable set of courses that can result in both a DH certificate and Public History credential within the timeline of the program. Due to these challenges, NULab faculty are currently reevaluating how to align our high expectations for DH training with the practical realities of a certificate program, which must exists alongside and in harmony with the primary curricular structures of humanities departments.

In addition to pressures on students, the popularity of the DH certificate among its first two cohorts of students has led to growing worry among departmental faculty that DH is driving down enrollments outside certificate program courses. We might be tempted toward market explanations ("we cannot dictate which courses students are interested in" or, less generously, "if our colleagues made their courses more enticing") but these are neither sufficient nor reflective. The NULab has created a certification that students perceive as necessary in a competitive job market, despite ambiguity about the role of DH in securing jobs (Risam 2013). Thus we have institutionalized a hierarchy of graduate course offerings that does privilege DH courses over others in the curriculum, in ways that partially reflect students' interests but partly reflect their anxieties. Moreover, the administration's vision for graduate education in the future clearly emphasizes digital humanities in ways that worry even NULab faculty. We cannot, in other words, entirely dismiss our colleagues' worries about how digital humanities, which belongs to no department in particular, has shifted the character and priorities of graduate programs in the particular departments of English and History.

In my presentation, then, I will think through what constitutes a successful DH graduate curriculum in an institutional culture of abundant top-down support and atrophying bottom-up enthusiasm. Can we structure robust DH training in ways that integrates with rather than competing with departmental training? Can a DH program be partner rather than usurper?

Miriam Posner

UCLA's Digital Humanities graduate certificate, founded in 2011, now enrolls 24 Ph.D. and master's degree students from across the university. Initially, the certificate was conceived as a means of providing an imprimatur for work that was already taking place at the graduate level. UCLA's humanities graduate students were already apprenticing on a wide range of faculty-led digital humanities projects, such as HyperCities and RomeLab, and developed a great deal of discipline-specific expertise through these experiences. The university, moreover, has a strong community of faculty DH practitioners and an established tradition of integrating graduate students into projects as collaborators. But as formal DH curricula grew at other institutions, UCLA graduate students and faculty began to feel that a formal credential might be useful to graduates as they entered the job market.

Even as students clamored for it, the introduction of an official graduate certificate has also had the effect of surfacing some challenges and dilemmas for graduate education in the digital humanities: how much of the graduate curriculum should be formal, and how much should come in the form of project work; how to provide the time- and resource-intensive instruction graduate students require; how to help students balance traditional dissertation work with digital work; how to accommodate the very distinct needs of Ph.D. students and professional master's degree students; and how to prepare graduate students for an unpredictable job market. Alexander Reid identified many of these dilemmas in a 2012 essay for *Debates in the Digital Humanities*, observing an uptick in digital humanities activity and arguing that we would soon witness a widespread shift in the education of scholars, toward an understanding of digital literacy as fundamental to graduate training (Reid 2012).

From the vantage of 2016, the picture seems less clear. Digital humanities continues to thrive as a field, but we have yet to see the searching, widespread reevaluation of graduate education that some observers expected. While a number of standout programs, such as the City University of New York's Graduate Center, the University of Virginia's Scholars' Lab, and the University of Victoria's Electronic Textual Cultures Laboratory, have seemed to forecast change on a larger scale, most humanities graduate programs still deal only gingerly, if at all, with digital technology.

The example of UCLA might help to illuminate some reasons for this very piecemeal rate of change. UCLA's graduate students, like most graduate students, are under enormous pressure and feel pulled in multiple directions by an erratic and whimsical job market. Assailed by advice to publish in top journals on the one hand, and to develop digital skills on the other, they often come to the DH program ready to perform a cost-benefit calculation about how this training will position them on the job market -- not exactly the spirit of embracing failure and creative experimentation that many DH experts advise (Ramsay 2010, Drucker 2009, Sample 2012). Devising a curriculum that makes sense for them in this climate, then, is constantly demanding and resource-intensive. Among the questions UCLA faculty has faced:

- Should a digital humanities program for graduate students emphasize collaborative scholarship, as many practitioners advise, or should students' DH work advance the individual dissertation?
- What is the program's responsibility toward preserving and archiving student digital work, and particularly digital dissertations?
- If a graduate DH program remains interdisciplinary, how can it assemble and retain the necessary core faculty to staff the program?
- How can an interdisciplinary program retain a "center of gravity" sufficient to enable graduate students to feel as though they are part of a community?
- How can a graduate DH program communicate its value to students' advisers, many of whom do not engage in digital work themselves?
- Given the highly individualized nature of dissertation-level work, how can graduate DH programs provide sufficient resources (staff time and technical assets) to help students advance their research meaningfully?
- How should a graduate program in DH balance the distinct needs of professional master's degree students (in UCLA's case, MLIS students) with Ph.D. students?

In this presentation, I will discuss the ways in which we at UCLA have attempted to develop a graduate curriculum that makes sense for a program that faces challenges familiar to most universities: lack of resources, little centralized support, and overtaxed faculty. I will also raise some questions about the sustainability of a digital humanities graduate curriculum without answering some searching and difficult questions about what a graduate program should be and do. Finally, I will propose some infrastructural and institutional solutions to help address some of the most pressing needs of graduate DH programs.

Maria Sachiko Cecire

At Bard College, we don't have a formal Digital Humanities center. Instead, we have an interdisciplinary curricular initiative and hub for faculty collaboration that we call Experimental Humanities (EH). DH scholarship at its most visible typically creates and employs digital tools to pursue project-based humanities research. While EH does some of this, our program was designed to align with the mission of our undergraduate-focused institution, and to be flexible enough to bring together faculty from diverse intellectual and personal backgrounds. We say that Experimental Humanities is Bard's liberal arts-driven answer to the Digital Humanities: it uses a network of courses and faculty-identified research clusters to variously interrogate how technology mediates what it means to be human. EH engages with media and technology forms from across historical periods, including our own, and combines experimental research methods with critical thinking about way media and technology function as a part of cultural, social, and political inquiry. We encourage the reconsideration of older media in light of today's technologies, and look ahead to the developments on the horizon.

The decision to name our program Experimental Humanities instead of Digital Humanities was grounded in the unique character and history of Bard College, which has developed an international reputation for its commitment to the arts, humanities, and the notion that access to a liberal arts education should be a fundamental human right. Although our primary campus in the Hudson Valley of New York serves just over 2000 students, Bard has a much wider reach that includes degree-granting programs in state prisons, early college programs at high schools that serve low-income youth in cities around the U.S., and in international university partnerships that bring liberal arts curricula to countries such as Kyrgyzstan, Palestine, and Russia. In this context of passionate liberal arts advocacy, we wanted a title that would leave room for DH but not exclude non-digital forms of artistic and scholarly production. As Wendy Chun has suggested, terminology that draws hard lines between "old" and "new" technologies runs the risk of excluding (or at least seeming to exclude) the lessons, theories, and knowledge of the past from the practice and study of "new" media. The notion of the "experimental" embraces both digitality and previous moments of technological change, invokes the practices of both the sciences and the arts, and has the kind of hands-on and countercultural associations that align with the Bard ethos.

We saw the rise of DH as an exciting opportunity to establish a program dedicated to reconsidering the methods and subjects of humanistic study in the light of changing material conditions. This is an ongoing project, and one that also allows us to continuously re-evaluate our pedagogical approaches: rethinking which tools and methods we use in the classroom and encouraging in-class reflection with our students about the relationships between who we are, what we study, and how we study it. For instance, when faced with N. Katherine Hayles's work on hyper and deep attention, students may put forward arguments for practicing deep attention in their coursework or advocate for a pedagogy that introduces more multimediated and hands-on content, thereby creating the opportunity for jointly designed assignments. Finally, in keeping with the concerns raised by #transformdh (see, for instance, Moya Z. Bailey's essay "All the Digital Humanists Are White, All the Nerds Are Men, but Some of Us Are Brave") we wanted to keep cultural critique and questions of inclusion central to what we do as humanities scholars.

With these ideals in mind, we set out to design a program that would be critical, inclusive, undergraduate-focused, and also able to participate in wider DH networks. In their essay about whether or how small liberal arts colleges might "do" DH, Bryan Alexander and Rebecca Frost Davis outline several challenges to establishing DH programs and centers at institutions like Bard. They note a lack of infrastructure to support major research projects, the difficulty of pulling together the human resources to do work that requires a wide range of skills, our limited access to graduate students that can sustain long-term research, and the pedagogical focus at SLACs. However, they argue that models that include curricular elements and partner with existing campus resources like library and IT can still be successful, developing proficiency in select project areas and sending students on to DH graduate programs.

Experimental Humanities does work closely with Library/IT and encourage faculty projects through training opportunities, the guidance of a Digital Projects Coordinator with a PhD in the humanities, and the support of a student Media Corps. But while several successful DH programs at other small liberal arts colleges have grown out of the library, like Occidental's Center for Digital Liberal Arts, or out of faculty research, as with Hamilton's Digital Humanities Initiative, EH was from its first imaginings a primarily curricular initiative, built on the three pillars of history, theory, and practice. All EH students take the core courses "A History of Experimentation" and "Introduction to Media," which EH faculty rotate teaching, and at least one practice-based course beyond the college arts requirement (this may include Computer Science or one of the visual, written, or performing arts). Students also take at least two more courses from the wide offering of EH-listed courses designed by faculty according to their research interests, and which are available each semester in fields from Music and Anthropology to Medieval Studies and Theater. These classes present the hands-on projects and research that they do at collective Share Events each semester.

At an administrative level, EH is a concentration (like a minor; Bard loves to have its own terminology for everything), which means that our students pair their coursework with a foundation in a major program of study, and that EH faculty also belong to a home program. All Bard students do yearlong senior projects, which allow our students to bring what they have learned in the concentration into conversation with their major discipline in a capstone project. We have seen senior projects that use topic modeling to analyze slave narratives, develop gaming apps with the potential to treat psychopathy, push the boundaries of traditional interview-based ethnography to consider the social implications of conversing via text message, delve into the history of the book and other media forms, and lead to immersive art installations in both digital and analog formats. Our students do not necessarily go into DH programs (though some do), but rather become curators, teachers, librarians, programmers, artists, and work for non-profits.

This kind of breadth makes Experimental Humanities sustainable even on a small campus, providing a hub for a range of interests and methods. Meanwhile, our regular course and event offerings give the program continuity and create opportunities for faculty and students to meet during the semester as a self-identifying community. We have also worked to bolster faculty research in recent years, beginning with the launch of our faculty-led, topic-based clusters in 2014. The clusters have been very successful in bringing faculty together to share and further their own research across disciplinary boundaries, and have yielded a number of new courses and given rise to a form of experimental symposia that bring together scholars, artists, practitioners, students, and community members around cluster topics such as "Sound" and "Surveillance." The clusters have become a model in the college for how to foster interdisciplinary collaboration that encourages both new research and student engagement. Other research models in EH include faculty-led humanities labs around individual projects, intensive student sessions with historical societies to create digital repositories and interfaces for the public to access its local history, and courses with embedded digital projects that allow faculty to work with undergraduates to build up layers of data each time that they teach the course.

EH is now in its fifth year, and coming to the end of a three-year grant from the Mellon Foundation. In my presentation I will discuss several of the challenges that we face at this crucial stage of transition. These include the ongoing struggle to define what we do and why it's useful to students, parents, and future employers, given our capacious title and mission; how to practically negotiate the need for disciplinary foundations in our students' and faculty's home programs and the invitation to experiment in EH courses and projects; how to move off of a major grant to become sustainable within the existing college structure (and thereby resist the kind of dependence on external grants that is contributing to the neoliberalization of the humanities, as outlined in the widely circulated "Dark Side of the Digital Humanities" papers); and how to better integrate the wider Bard network of underserved high school and undergraduate students both in the US and around the world into the work that we do.

Bibliography

Alexander, B., and Davis, R. F. (2012). "Should Liberal Arts Campuses Do Digital Humanities? Process and Products in the Small College World," in *Debates in the Digital Humanities*, ed. Matthew K. Gold http://dhdebates.gc.cuny.edu/debates/text/89

Bailey, M. Z. (2011) "All the Digital Humanists Are White, All the Nerds Are Men, but Some of Us Are Brave," *Journal of Digital Humanities* 1.1 http://journalofdigitalhumanities.org/1-

1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/

Burdick, A., Drucker, J., Presner, T., Schnapp, J., and Lunenfeld, P. (2012) *Digital_Humanities*. Cambridge, MA: The MIT Press.

centerNet (n.d.). "centerNet: An international network of digital humanities centers". Web. http://dhcenternet.org/centers

Chun, W. H. K. (2006) "Introduction: Did Somebody Say New Media?" in *New Media, Old Media: A History and Theory Reader*, ed. Wendy Chun and Thomas Keenan (New York: Routledge), 1-10.

Chun, W. H. K., Grusin, R., Jagoda, P., and Raley, R. (2016)"The Dark Side of the Digital Humanities," in *Debates in the Digital Humanities*, ed. Matthew K. Gold . http://dhdebates.gc.cuny.edu/debates/text/89

Clement, T. (2015) "The Information Science Question in DH Feminism." *Digital Humanities Quarterly*, vol. 9, no. 2.

Drucker, J. (2012) "Humanistic Theory and Digital Scholarship." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, Minneapolis: University Of Minnesota Press, pp. 85 – 95.

Drucker, J. (2009). SpecLab: *Digital Aesthetics and Projects in Speculative Computing*. Chicago: University of Chicago Press.

Gold, M. and Klein, L., eds. (2016). *Debates in Digital Humanities*. Minneapolis: University
of Minnesota Press.

Hayles, N. K. (2012) *How We Think: Digital Media and Contemporary Technogenesis*. Chicago; London: University Of Chicago Press.

Hayles, N.K. (2007). "Hyper and Deep Attention: The Generational Divide in Cognitive Modes," *MLA Profession* (2007): 187–199.

Liu, A. (2016) "Drafts for Against the Cultural Singularity" Alan Liu. 2 May 2016.

Modern Language Association. (n.d.) "Digital Pedagogy in the Humanities: Concepts, Models, and Experiments". MLA Commons,. Web. https://digitalpedagogy.commons.mla.org/

Off the Tracks (2011). Collaborators' Bill of Rights. Media Commons Press. Web. http://mcpress.media-commons.org/

Ramsay, S. (2010). "The Hermeneutics of Screwing Around; or What You Do with a Million Books." (2010). Unpublished presentation delivered at Brown University, Providence, RI 17 (2010): n. Pag. Web. 17 Oct. 2012.

Rector and Visitors of the University of Virginia. (2017). DH@UVA. Web. http://dh.virginia.edu

Rector and Visitors of the University of Virginia. (2008) "Sciences, Humanities and Arts Networked Technology Initiatives," SHANTI at the University of Virginia. . Web. http://shanti.virginia.edu/

Reid, A. (2012)."Graduate Education and the Ethics of the Digital Humanities." *Debates in the Digital Humanitie*s. Ann Arbor: University of Michigan, Print.

Risam, R. (2013) "Where Have All the DH Jobs Gone?" Roopika Risam 15 Sept. 2013. Web. 1 Nov. 2016.

Sample, M. (2012) . "Notes towards a Deformed Humanities." samplereality. N.p., 2 May 2012. Web. 29 Oct. 2016.

Svensson, P. (2012) "Envisioning the Digital Humanities" *Digital Humanities Quarterly*, vol. 6, no. 1.

Svensson, P. (2009) "Humanities Computing as Digital Humanities" *Digital Humanities Quarterly*, vol. 3, no. 3.

The Leadership Alliance (2016). "Mellon Initiative" . The Leadership Alliance. Web. http://www.theleadershipalliance.org/programs/summer-research/mellon-initiative

UCLA Center for Digital Humanities. (n.d.) "UCLA Digital Humanities". Univeristy of Claifornia Los Angeles. Web. http://www.cdh.ucla.edu/

University of Virginia. (n.d.) Institute for Advanced Technology in the Humanities. Web. http://iath.virginia.edu/

University of Virginia College and Graduate School of Arts and Sciences (n.d.) "Learning Design Technology, Arts and Sciences". University of Virginia.. Web.
http://learningdesign.as.virginia.edu/

University of Virginia Library. (n.d.) Scholars' Lab,. Web. http://scholarslab.org/

Verhoeven, D. (2016) "As Luck Would Have It." *Feminist Media Histories*, vol. 2., no. 1. pp. 7–28

Washington and Lee Digital Humanities Action Team (n.d.). "DH @ W&L". Washington and Lee University. Web. http://digitalhumanities.wlu.edu/

Digital Religion – Digital Theology

Claire Clivaz
claire.clivaz@sib.swiss
Swiss Institute of Bioinformatics, Vital-IT, Switzerland

Emily S. Clark
clarke2@gonzaga.edu
Gonzaga University, United States of America

Katherine M. Faull
faull@bucknell.edu
Bucknell University, United States of America

Paul Dilley
paul-dilley@uiowa.edu
Iowa University, United States of America

Rachel McBride-Lindsey
lindseyrm@slu.edu
St Louis University, United States of America

Peter Phillips
p.m.phillips@durham.ac.uk
CODEC Research Center
Durham University, United Kingdom

Introduction

Scholarly discourse evaluating the digital turn in biblical and religious studies is at an early stage in its development, as attested to by the creation of two new book series in 2016: *Introduction to Digital Humanities: Religion* (IDH, de Gruyter), and *Digital Biblical Studies* (DBS, Brill). Previously, Heidi Campbell published an overview of the topic (Campbell 2013), developed in further publications (Campbell-Althenhofen 2015, Campbell-Garner 2016). In a recent overview, Carrie Schroeder develops two central questions on the topic: “what does it mean for Biblical Studies to be marginal to the Digital Humanities when DH is a field positioning itself as transformative for the humanities? How can our expertise in Biblical Studies influence and shape Digital Humanities for the better?” (Schroeder 2016). Using her field, Coptic studies, as an example she shows that the particular skills and needs of a marginal field within a marginal field can be a strong driver in DH.

Consequently, and for the first time at a DH meeting, this ninety-minute panel session asks what is the impact of the digital turn on religious studies and theology, and to what extent these somewhat marginal fields can bring something specific to the big DH tent. They particularly focus on textuality and on the symbolic impact of the “book” as attested to in the expression, “religions of the book,” coined in a programmatic lecture given in 1870 by F. Max Müller (2010). The symbolic, Western impact of books and writing was amplified by this notion, born at the time when the legal status of printed texts and authorship was completely secured in Western culture (Clivaz 2012).

For centuries, “books were perceived as a ‘wide angle’ from which it was possible for everything to be observed, related to, and perhaps even decided” (Carrière-Eco 2009). The panel will consequently consider the hypothesis that the DH have been deeply influenced by this fascination with textuality and books during the first decades of their development; while keeping “the discourse of written texts” as a central pillar to the discussion according to the words of Roberto Busa, a foundational DH figure (Busa 2004). Busa’s relationship to Biblical and religious materials has played a role in his approach to the computing field, as Jones point out (Jones 2016). The double impact of the book and the notion of “religions of the book”, successful in Western culture since the 19th century, provides an opening to understanding why DH in religious fields is still so focused on textuality. Indeed, when we collect examples of DH studies in diverse religious fields, we are unsurprisingly faced with very textual DH (Clivaz et al. 2016e). This observation strengthens the necessity for religions in DH to consider the multimodal and multicultural turn provoked by digital culture.

With these different questions in mind, five panelists will participate in the presentations (sixty minutes in total) and a thirty-minute panel discussion that will be moderated by Claire Clivaz representing the Swiss Institute of Bioinformatics, Vital-IT (Lausanne, CH). The following five speakers have agreed to participate and to discuss the general topic from the perspectives of their own research projects. In alphabetical order:

A Neophyte Proselytizes for Digital Humanities Pedagogy

Emily S. Clark

This presentation explores the ways in which Digital Humanities can enhance a Religious Studies classroom by focusing on two assignments that ask new questions of traditional course materials. The first is a project that was the culmination of a month's work collaboratively amongst a class of 25 students with a database platform (Omeka). This project entailed the digitization of archival photographs of a Native American community from 1916, along with the reading of Jesuit mission material (Clark et al. 2016). The second is an assignment that took two class periods and introduced students to data visualization (Voyant). This assignment introduced students to the differences between close reading and distant reading, along with practicing both on excerpts from Jesuit mission documents (Mentrak – Bucko, 2016).

Topic Modeling the Bible

Paul Dilley

The talk will present the first full-scale topic model of the Bible and related literature in four different languages: Greek, Latin, Syriac and English. It will discuss both technical aspects of the process (e.g., the use or not of lemmatization; retention or removal of function words; optimal number of topics), as well as what we gain from comparing topic models of the same corpus translated into different languages. The presentation will focus on the interpretive gains and losses involved in topic modeling, one of the richest strategies of distant reading to the Bible which has been the subject of centuries of minute examination of the close reading tradition which Moretti has pointedly labeled a “theological exercise” (Moretti 2013).

Digital Lives: Reading Moravian Memoirs in the Age of the Internet

Katherine M. Faull

An international collaborative research project (USA, Sweden, Germany) is developing a digital platform for the investigation of the metadata and text of Moravian memoirs, composed since the mid-18th century by members of the Moravian Church to be read at their funeral (over 65,000 memoirs, housed in Germany and the US, Faull 1997). Less than 10% of the earliest manuscripts have been published. The developing digital interface allows for geospatial and chronological visualization of author’s birth and death place (Haskins 2007). This paper will investigate the intersection of the digital, the autobiographical, and the sacred in the age of the internet. How can the act of reading the lives of thousands of Moravians also be understood as an act of reconstituting the “invisible church” ? (van Dijk, 2007; Eakin 2014).

Material Religions in a Digital World

Rachel McBride-Lindsey

For much of the modern era, religion and theology have been intertwined in a decidedly material world. Over the last several decades, students of religion have begun to carve out intellectual headroom for an approach to material culture that recognizes objects and images as generative sources of theological inquiry and religious practice. Cultural institutions can be an effective tool for inviting researchers and the public into physical spaces and into contact with deeper dimensions of the material world. At the same time, these very contributions work against methodological gains in the study of material culture. Rachel McBride-Lindsay's presentation starts with this tension and draws from pedagogical attempts to incorporate digital platforms into projects anchored in the study of objects.

Exploring developmental patterns within Digital Theology Research within the Digital Humanities

Peter Phillips

Campbell and Altenhofen (2015) explore four waves in digital research development in theology and religion back into the late twentieth century. Their wave pattern picks up both historical and technological trends and patterns in research. However, a three wave theory dominates discussion within introductions to the Digital Humanities, discussed by David Berry (2011) and in the *Digital Humanities Manifesto 2.0*. It tends to reflect modes of research, or groups of methodologies used in research rather than time periods. Reflecting on CODEC's own experience of Digital Theology in association with a range of other scholars, this paper will assess whether too many waves are a problem in our methodological theorizing.

Bibliography

Berry, D (2011), "The computational turn: thinking about the Digital Humanities", *Culture Machine* 12, 1–22.

Busa, R. (2004), "Foreword: Perspectives on the Digital Humanities", in S. Schreibman, R. Siemens, J. Unsworth (ed.), *A Companion to Digital Humanities*, Oxford: Blackwell, http://www.digitalhumanities.org/companion/ (Accessed 27 March 2017).

Campbell, H.A (2013) (ed.), *Digital Religion. Understanding religious practice in new media worlds*, London/ New York : Routledge.

Campbell, H.A. and Altenhofen, B (2015), "Methodological Challenges, Innovations and Growing Pains in Digital Religion Research", in *Digital Methodologies in the Sociology of Religion*, S. Cheruvallil-Contractor – S. Shakkour (eds.), Blumsburry Publishing, Kindle edition.

Campbell, H.A. and Garner, S. (2016), *Networked theology. Negotiating faith in digital culture*, Grand Rapids, MA : Baker Academy.

Carrière, J-.C. and Eco, U. (2009), *N'espérez pas vous débarrasser des livres*, Paris, Seuil.

Clark, E.S. et al. (2016), *Digital Jesuits and Ignatian Pedagogy*, King Island Collection, Jesuit Oregon Province Archives, Gonzaga University, http://as-dh.gonzaga.edu/omeka/ (Accessed 27 March 2017)

Clivaz, C. (2012a), "Homer and the New Testament as 'Multitexts' in the Digital Age ?", *SRC* 3/3, 1-15 ; http://src-online.ca/index.php/src/article/view/97 (Accessed 27 March 2017).

Clivaz et al. (eds.) (2016), *Digital Humanities in Jewish, Christian and Arabic traditions*, special issue *JRMDC* 5 (2016/1), https://www.jrmdc.com/journal/issue/view/9 (Accessed 27 March 2017).

Von Dijk, J. (2007), *Mediated Memories in the Digital Age*. Stanford: Stanford UP.

Eakin, P. J. (2014), "Autobiography as Cosmogram", *Storyworlds: A Journal of Narrative Studies* 6/1: 21-43.

Faull, K. M. (1997), ed. and trans., *Moravian Women's Memoirs: Their Related Lives, 1750–1820*, Syracuse, NY: Syracuse University Press.

Haskins, E. (2007), "Between Archive and Participation: Public Memory in a Digital Age", *Rhetoric Society Quarterly* 37/4, 401–422.

Jones, S. (2016), *Roberto Busa, S. J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*, Routledge Press, London

Mentrak, T. and Bucko, R.A. (2016), *Jesuit Relations and Allied Documents 1610 to 1791*, http://moses.creighton.edu/kripke/jesuitrelations/ (Accessed 27 March 2017).

Müller, F.M. (2010), "Second Lecture Delivered at the Royal Institution, February 26, 1870", in F. M. Müller, *Introduction to the science of religion. Four lectures delivered at the Royal Institution*, Seneca Falls, NY: Wilson Press, 52–82 (Original lectures February & May 1870).

Moretti, F. (2013), *Distant Reading*, Verso, London, New York.

Schroeder, C.T. (2016), "The Digital Humanities as Cultural Capital: Implications for Biblical and Religious Studies", *Journal of Religion, Media and Digital Culture* 5(1), 21–49, <http://www.jrmdc.com/journal/issue/view/9> (Accessed 27 March 2017).

Queer Infrastructures: Digital Intimacies, Spaces, Affordances and Collaboration

T.L. Cowan
tlcowan1@gmail.com
Yale University, United States of America

Jasmine Rault
jasrault@gmail.com
The New School, United States of America

Dayna McLeod
dayna.mcleod@concordia.ca
Concordia University, Canada

Carina Guzmán
cartogeosapiens@gmail.com
Concordia University, Canada

This panel centers queer epistemological formations as they structure the study of digital phenomena. Each of the papers mobilizes the methodological foundations of what Kara Keeling has called a “Queer OS” (2014) -- or an operating system in which “the historical, sociocultural, conceptual phenomena that currently shape our realities in deep and profound ways, such as race, gender, class, citizenship, and ability (to name those among the most active in the United States today), [are understood as] mutually constitutive with sexuality and with media and information technologies, thereby making it impossible to think any of them in isolation” (152). The panelists address core questions including how digital infrastructures mirror and/or model queer and feminist approaches to scholarship, social and artistic practice; what are some possible aesthetic/political retoolings for digital technologies originally designed to surveille and/or maximize female bodies as reproductive vessels; how do social media platforms shift the organizing of live performance cultures?; what are some of the genealogies for contemporary queer digital culture and design?; how can we mobilize theories of interiors and architecture to help us think about the design of online spaces?; what can we learn from long histories of queer social, erotic, political and artistic life that can help us to build online platforms that resist calls to full public access?

These four papers contribute to the larger conversation at the nexus of Digital Humanities, Digital Media Studies, and Theories and Practices of Transgender, Feminist and Queer Cultural Production. Here we bring together the practice of performance art that merges digital technological invention to the long history of feminist body art, and the often invasive politics of digital surveillance and control of women’s bodies, along with the introduction of “Transmedial Drag” as a set of methods for analyzing the digital transfer and potentiality of live performance art, analog queer networks and spaces. We bring this together with a discussion of the development of digital and analog infrastructures for minoritarian spaces and scenes in Mexico City and a material analysis of many of the tropological framings of digital culture--screens, codes and filters--which have been centrally important to the field of feminist design and architectural studies.

Transmedial Drag: Transgender, Feminist and Queer Performance Art in the Age of Digital Reproduction

T.L. Cowan

One of the central foci of Digital Humanities scholarship has been the production of online repositories of previously inaccessible archival materials. Often the project of producing an online repository is generated by the assumption that open access to materials that were once held behind the walls of institutional protocols or hidden in personal collections, is a good thing both for the research community that may want access and for the materials themselves and the lives and events that they assemble. However, as scholars invested in the project of decolonization--including Mukurtu and Local Context co-founders Jane Anderson and Kim Christen--have show us, making openly accessible all of the archival materials collected about a particular community, will very likely violate the cultural protocols of that community. Indeed, the entitlement required to imagine that all materials about a particular community should be made accessible to people who are not part of that community is a perspective with long colonial and imperialist genealogies.

In this paper, I present a range of speculative-pragmatic methods that my collaborators and I have used towards the development of The Cabaret Commons: An Online Archive and Anecdotal Encyclopedia for Trans- Feminist and Queer Artists, Activists and Audiences. Over the course of the past 5 years of community-engaged research, we have worked both to assemble materials from live performance scenes in Mexico City, New York City, Montreal and Toronto to make publicly accessible for researchers and artists alike, and to design the Cabaret Commons as an online platform structured by community protocols, filters, ethics, and aesthetics of the stages and performance spaces that structure the social-cultural scenes that produce these materials.

After years of working with the idea of the Cabaret Commons, I named our method “Transmedial Drag,” and have elaborated how we work in the mode of transmedial drag along with co-investigator Jasmine Rault, and collaborators Dayna McLeod, Carina Guzmán and Robyn Overstreet as well as many performance artists with whose work we are in conversation.

“Transmedial drag” is the method of study, knowledge production and citational practice involved in moving across media/mediums (for example, from live performance, to video documentation, to digital archive to online platform), which creates a sort of pastiche of the ‘original,’ denaturalizing its status as ‘originary’ and teaching us something new about the excesses and limitations of each media form. Riffing on Judith Butler’s game-changing theorization of drag as a performance that, “[i]n imitating gender ...reveals the imitative structure of gender itself--as well as its contingency” (Gender Trouble 138). Or, thinking through Benjamin’s anxieties about transmedial transfer in the early 20th Century, it’s what we might call *the work of queer performance art in the age of digital reproduction…!* Thus we practice a process-heavy collaborative protocol that seeks to “simulate,” not just the “aura” of live performance, but is also structured by locally-specific, community-based social and political ethics as manifested in minoritized cabaret scenes. Transmedial drag compels us to reckon with the cultural and ethical limitations of digital accessibility in the face of technical possibility (cf. Robertson 2016).

The decolonizing, community-collaboration and accountability methods for the digital transmission of traditional Indigenous knowledges, practiced for example by the projects Mukurtu and Local Contexts, provide a generative model for negotiating this paradox of intimate privacies and networked cultural histories and memories. Indeed, Indigenous studies has pointed out the colonial and Westernexpansionist logics and impulses underlying the push to open-access and digitization, and have developed networked information management and archiving systems that follow indigenous cultural protocols, prioritizing privacy and multi-tiered user-generated-access levels.

What we're calling transmedial drag refers to the troubling excess that remains in even the most 'successful' transition from one medium or digital form to another – those elements that are incompatible to the form/medium as it currently exists and compel us to ask whether the form/medium exists *because of* these constitutive exclusions, or on the condition that certain elements remain excessive/external. Indeed, what remains excessive, incompatible and bit of a drag on the seductive techno-cultural possibilities of openness, are things like labour, consent, privacy, the agency and will of research-collaborators. For those of us trained in gender, sexuality, critical race, post and decolonial research, it becomes clear that our scholarly fields and practices of digital media and digital humanities can be sustained by reproducing a familiar set of constitutive exclusions -- those skills, values, techniques and priorities cultivated by and for the survivance of women, people of colour, indigenous people, trans people, queers, etc.

Transmedial drag is indebted to Elizabeth Freeman's work on 'temporal drag' and Heather Love's on "feeling backward" -- or the ways that queerness is often experienced, expressed or interpreted as an attachment to the past, as a stall in personal or social development, a drag on cultural-temporal progress narratives. We are also indebted to Kadji Amin's work on transgender temporalities, which allow us to understand the ways that "transgender experiences are constituted by yet exceed normative temporalities" and the necessity of incorporating asynchronic temporalities "in order to do justice to the complex ways in which people inhabit gender variance" (220).

With Love, Freeman and Amin, our digital research-creation project is oriented to those people, scenes, cultures, affects and knowledges cast as a drag on the forward-momentum- digital-entitlement fantasies of unfettered access to everything, everyone, everywhere -- fantasies that resonate so strongly with the gender, sexual and racial histories of colonial-modernity that we know too well. Thus in true queer epistemological form, "transmedial drag" is both a method of study modeled on queer performance and theories of performativity (i.e. Butler) and a conceptual framework through which to make sense of elements of cultural life and work that are generally unintelligible (nonsense).

Like the deep cultural understandings, consultation, collaboration and re-engineering that led to the development of Mukurtu and Local Contexts, we imagine culturally-sensitive and specific software that prioritizes highly nuanced **cultural** logics, protocols and specialized knowledges in the development of **computational** logics, protocols and specialized tools, towards what Kara Keeling calls a "Queer OS/operating system". Keeling's QUEER OS is an operating system in which the "historical, sociocultural, conceptual phenomena that currently shape our realities in deep and profound ways, such as race, gender, class, citizenship, and ability (to name those among the most active in the United States today), [are understood as] mutually constitutive with sexuality and with media and information technologies, thereby making it impossible to think any of them in isolation" (152).

Indeed, transmedial drag denaturalizes the constitutive exclusions of most of our media and information technologies; lets us both recognize their limits and imagine their necessary transformations; **and** puts into relief some of the dynamics of cabaret that continue to sustain trans- feminist and queer scenes and lives **even within** media environments that are designed around their impossibility and imperceptibility.

In this moment when new digital archives of minority and subcultural scenes seem compelled to reproduce perilous Web 2.0 logics of unbridled open-access -- based in modern-colonial entitlement fantasies of absolute accessibility -- the Cabaret Commons is using trans feminist and queer relational epistemologies towards the decolonizing project of ethical collective platform protocols for sensitive cultural memory, heritage projects and accountable digital design. This is part of a larger epistemological-transformational project to decolonize both trans- feminist and queer practices, and the dominant digital cultural practices that assumes that all networks are networked publics leading us to theorize what we're calling networked privates.

The hope that we have for the Cabaret Commons is that it will bring the activated characteristics of cabaret performance, as well as other grassroots and politically-engaged live performance (like street performance, marches and protest arts -- what Mexican artist/activist Jesusa Rodríguez calls "mass cabaret") along with their translocal trans feminist and queer scenes, ethics, politics, social and sexual lives to bear on digital archiving infrastructures.

Screens, Codes, Filters: sapphic modernist design genealogies for queer digital culture

Jasmine Rault

What can early twentieth century modernist architecture and design teach us about contemporary decolonizing, feminist, queer and anti-racist practices and protocols for networked digital architecture? Euro-Atlantic modernist architecture was driven by ideals of internationally standardized open communication – beyond the mass rail, steamboat and automobile distribution of low-cost print media and photographic reproductions towards material changes like the open floor plan, strip windows and glass walls, unimpeded visual access to interior and exterior space, eliminating walls and structural as well as decorative or symbolic obstructions to complete open access. There are some striking ideological continuities between these modernist architectural ideals (and aesthetics) and contemporary Euro-Atlantic values of unbridled digitization, designing global information networks for unobstructed open access – and these continuities need to be understood within a context of modern-colonial regimes of gender, sexuality and race. My presentation focuses on early twentieth century women's queer interventions into modernist architectural and design ideals as a genealogy to contextualize, better understand and support recent innovations in decolonizing, queer, feminist and anti-racist designs for networked digital practices, archives and spaces. I draw from my research on Eileen Gray – which explores some of the intersections between histories of communication, European architectural modernity and sapphic modernity, or the cultural history of female sexual dissidence – and my collaborative research on "the Cabaret Commons," designing a networked digital archive for trans feminist and queer performance artists, activists and audiences.

My research on Gray, and her contemporaries in queer interior designs, shows that modernist architecture was invested in creating not only new buildings and living spaces, but new bodies and subjects. As such, modernist architecture was contributing to the modern production and regulation of sexuality, race and gender. Moreover, from around 1900 to 1935, modernist architecture was increasingly committed to communicative clarity – or immersive living spaces of total communication – and several sexually dissident female artists, writers, interior designers and architects, like Eileen Gray, worked to interrupt this clarity. Indeed, women in the US, Canada, Western Europe and the UK worked on domestic interior design to work out new possibilities for gender and non-heterosexualities.

I focus on the creative extent to which these possibilities depended on codes, screens, filters – as interpretive and communicative strategies, but also as design materials. That is, the aesthetic, social and cultural phenomena of female gender and sexual queerness was enabled at the start of the 20c through complex claims to privacy. My presentation draws connections to my current book project, co-written with T.L. Cowan, provisionally entitled, *Checking In: Feminist Labour in Networked Publics,* which takes up the ways that similar claims to mediated privacy are emerging as central concerns in the design of very different spaces – online digital designs for networked trans-feminist and queer archives and social memory.

We are currently working with the data set of the digital (digitized) archive of the Meow Mix Cabaret, a show and dance party for "bent girls and their buddies" which ran in Montreal from 1997-2012. Technically, we have been granted permission by the copyright holders to make this archive public through an online platform. However, through interviews with individual performers who have appeared on the Meow Mix stage over the years, we've learned that a majority of the artists are not interested in reproducing or getting involved in the distribution of a full-run, open-access digital archive of their materials – for many reasons: including low or degraded quality of the video and images; the unpolished/amateur aesthetic of their work at an earlier stage of their careers; the fact that they did the performance *for their friends* or for a particular event and do not want broad circulation that will leave a digital trace; gender, sex, sexuality, body shape and size transitions; nudity; and the potential hazards of being associated with trans- feminist and queer scenes. However, many artists are interested in *some* of their work being available online for *some* people. And thus by working through what we are calling a speculative/pragmatic method of transmedial drag, we are designing an online space in which this kind of mediation will be possible.

In this presentation, I suggest that understanding the architectural innovations of Gray and her contemporaries, who built on the premise of mediated privacy, on filtered access, offers us a unique genealogy to studying and designing digital architecture that respond to similar needs and desires for online spaces through which minoritized subjects push back against the dominant pressure for full publicity, for the full availability and open access to our online selves, socialities and intimacies. The metaphor of architecture is often used in reference to online built environments but rarely is the metaphor pushed to a point of usefulness. Danah Boyd has argued that "what it means to be public or private is quickly changing before our eyes and we lack the language, social norms, and structures to handle it" (2007). Indeed, queers, people of color, indigenous people, trans folks, and disabled people are hacking popular social media corporate sites like Facebook, Twitter, Instagram, Snapchat, etc. to build differentiating filters in order to choose which *selves* to "out" in which contexts, and organizations like the Feminist Technology Network (FemTechNet), the Center for Solutions to Online Violence (CSOV), FemBot, Crash Override (Zoe Quinn & Anita Sarkesian) are working not only to build these infrastructures but also to communicate the differential consequences of total publicity, open access, or the digital 'open plan.' That is, we can see the development of networked 'privates' rather than the fetishization of 'networked publics' and we are working on collective action towards, what we might have learned from Gray and her contemporaries: techniques of mediated privacy.

Queering the Vaginal Canal: Make Art Here

Dayna McLeod

My presentation will examine how feminist performance artists have employed explicit performance-based practices in their work, specifically by using their vaginal cavity as a site of art production. I will compare these works, practices, and methodologies with my own, specifically citing **Uterine Concert Hall**, a vaginal media work that features my body as a concert venue. Equipped with a 54 kHz internal speaker (Babypod), my vaginal canal acts as the stage with my cervix as the proscenium, for the audience of my uterus. A live DJ pumps sound directly into me via 6-foot cable that reaches from their booth. My vaginal canal is the scene of the performance and the instrument of its production. I can feel the DJ's varying frequencies, pitch shifts, and throbbing tones while external concertgoers are invited to eavesdrop via stethoscope, on the faint echoes of the recital through the very flesh of my body. Like showing up to a concert and listening from outside, this piece purposefully excludes external listeners while engaging with explicit performance-based production practices, and feminist art practices of intimacy.

In *The Explicit Body in Performance*, Rebecca Schneider identifies Annie Sprinkle's cervix, viewed similarly through a medical device (a speculum) in *Public Cervix Announcement* (1990-1992), as a "theoretical third eye," a countergaze that looks back at the viewer who is gazing at Sprinkle's cervix (55). In **Uterine Concert Hall**, the viewer's gaze and concentrated attention is focused on me and on the message my body embodies for their speculation: the spectator sees, hears, and touches my body while I absorb both the concert through my vagina and their expectations. My "theoretical third eye" listens. Schneider also notes that "any body bearing female markings is automatically shadowed by the history of that body's signification" (20), and the invasive imagined construction of that body. For an external audience to **Uterine Concert Hall**, my uterus-as-intended-audience - cervix-as-proscenium - vagina-as-stage, all become a singular blank screen of possibilities for viewers to project their fantasies onto: to imagine its appearance and construction, and imagine what sounds might emit from my flesh. How might my body distort or otherwise enhance the anticipated DJ set? (How) does sound transform the body (my body)? By the time the viewer arrives at my side where they are handed a stethoscope to eavesdrop into/onto/through my flesh, their anticipation and fantasy is confronted by the very real presence of the mise en scene. Here they are featured in the role of pseudo-physician, which is complicated by issues of consent, interior/exterior bodily access, and their demands of the body (my body) as sound medium.

This outsider, exclusionary status that the viewer is assigned also comes with a controlling, medicalized, obstetrical gaze. In Canada and the United States, we employ much monitoring of women's bodies through our expectations of how women look, behave, act, and feel. These expectations are pathologized, reinforced, and legitimized through medicalized surveillance and control. Invasive and non-invasive examinations and procedures like ultrasounds, transvaginal probes, and visualizing and monitoring technologies are normalized for bodies marked female who are evaluated by a normalized and medicalized gaze in relationship to potentially housing or not housing a fetus (Balsamo). "Protection of the fetus is often offered as a commonsensical, and, hence, ideological rationale for intervention into a woman's pregnancy, either through the actual application of invasive technologies or through the exercise of technologies of social monitoring and surveillance" (Clarke and Olesen). **Uterine Concert Hall** is not a place for babies, fetuses, or heteronormative determinism. However, it is a site that questions these cultural assumptions of women's bodies, of what we expect from bodies marked female, and why we think we have the right to make any kinds of demands on these bodies in the first place. This project does this by using digital technologies and affordances to interrupt their intended functions (i.e. playing music for a uterus-bond fetus from an adjacent vaginal canal) that contribute to the medicalized surveillance and control culture of women's bodies.

In this panel presentation, I am also interested in discussing **Uterine Concert Hall** in its queering of the uterus as a viable physical space, the performing body, and the artist-audience exchange. Further, I am interested in putting **Uterine Concert Hall** in conversation with the works and practices of artists who have similarly used vaginal and uterine space *as* physical space as the site of their works' production, like Annie Sprinkle's *Public Cervix Announcement* (1990-1992), Carolee Schneemann's *Interior Scroll* (1975), and Casey Jenkins' *Vaginal Knitting* (2013).

"How many lesbians does it take to flyer for a party?". The impact of shifting digital landscapes on queer women's nightlife organizing in Mexico City since 2005.

Carina Guzmán

In this paper, I open an inquiry on how the evolution of the digital landscape has brought shifts in the strategies of queer women's nightlife organizing in Mexico City since 2005, and what the possible implications of these shifts are.

Following Cowan and Rault's concept of transmedial drag, I ask if and how social media platforms and media such as memes, used as organizing/political tools, can possibly take the shape of/mimic/reflect the queer world-making of the social scenes they are being used to create, especially when contrasted with the resources they have replaced such as e-mails or flyers, which necessarily imply physical contact between organizers and the queer crowd. I pay special attention to independently organized events held outside the circuit of commercial gay nightlife venues; dance parties and cabaret revues in adapted spaces such as *azoteas* and *casas de cultura*. This leads me to consider the intersection of labour, material culture, digital culture, and the sexual politics of place-making. At the same time, I consider the shift within digital culture that has happened with the emergence of social media as an organizing tool, and memes as political statements.

In my doctoral research in the Communication Studies program at Concordia University I have established, partially through an autoethnographic narrative, that independent and capitalist-alternative lesbian and queer women's nightlife organizing in Mexico City is a political project that responds to issues of economic gender disparity and spatial justice.

The small lesbian collective *Meras efímeras* was formed in 2005 by a group of friends I was a part of to organize nightlife alternatives for queer women in Mexico City. At the time, the main nighttime recreational events for queer women were "ladies' nights" at gay men's bars and clubs. Though inexpensive and conveniently located in the city's gay ghetto, these were not especially well organized; patrons could often find poor service, strip shows that made some feel uncomfortable and unclean facilities. It was evident that they were held to create a niche clientele on slow weeknights.

Meras efímeras contended that within the capitalist logic of commercial nightlife, a queer women's crowd could not "compete" with an audience of gay men, making it virtually impossible to find women-oriented or women-welcoming queer events on the weekend. We, thus, understood that independently organizing nightlife implied taking a political stand on the economic disparity between men and women, and the issue of spatial justice for queer women in the urban nightscape. This also implied we could not expect to be paid for our work. And, as we were also unable to pay rent at a club, we had to physically adapt spaces or negotiate agreements at other types of venues outside of the gay ghetto.

Today's use of memes to make political statements, as well as the use of platforms such as Facebook event pages, Twitter and Instagram as an event organizing tools contrasts sharply with what it was like to call on a queer crowd for a political/social event around 2005. Our main digital resource was a Yahoo! Groups mailing list to which we would manually add e-mails requested at the door of our events. We also advertised in the local LGBT free weeklies. But, the first point of contact were flyers we'd drop off at businesses or hand-out outside "ladies' nights" events.

In this paper I establish that the simultaneous use of paper-based and e-mail group based strategies *Meras efímeras* employed around 2005 constitute a pre-social media-as-we-know-and-use-it-today organizing landscape that straddled material and digital culture. While this strategy was partially material and partially digital, it necessarily implied exchanges made physically; handing out of the flyer on the street, requesting and giving an e-mail at the door of an event. So, this paper also asks what are the stakes in queer women's nightlife organizing, as a political project, in the transfer from organizing strategies that implied a physical presence to current social media resources. Moreover, I ask if resources such as memes, Facebook event pages, Tweets or Instagram are capable of an effective transmedial drag; can they properly represent queer women's nightlife political world-views?

Bibliography

Amin, K. (2014). "Temporality." *Transgender Studies Quarterly* 1 (1-2): 219–22.

Balsamo, A. M. (1995) *Technologies of the Gendered Body: Reading Cyborg Women.* Durham NC: Duke University Press.

Butler, J. (1998). *Gender Trouble: Feminism and the Subversion of Identity*. 1st ed. New York: Routledge.

Clarke, A. E., Olesen, V. (2013) *Revisioning Women, Health and Healing: Feminist, Cultural and Technoscience Perspectives.* London and New York: Routledge

Freeman, E. (2010). *Time Binds: Queer Temporalities, Queer Histories*. Durham N.C.: Duke University Press Books.

Keeling, K. (2014). "Queer OS." *Cinema Journal* 53 (2): 152–57.

Love, H. (2007). *Feeling Backward: Loss and the Politics of Queer History*. Cambridge, Mass: Harvard University Press.

Robertson, T. (2016). "Digitization: Just Because You Can, Doesn't Mean You Should." *Tara Robertson*. March 20. http://tararobertson.ca/2016/oob/

Schneider, R. (1997) *The Explicit Body in Performance.* London and New York: Routledge.

High-resolution musicology: Capturing and encoding source detail for medieval music

Julia Craig-McFeely
julia.craig-mcfeely@music.ox.ac.uk
University of Oxford, United Kingdom

Karen Desmond
kdesmond@brandeis.edu
Brandeis University, United States of America

Ben Florin
benjamin.florin@bc.edu
Boston College, United States of America

Andrew Hankinson
andrew.hankinson@music.ox.ac.uk
University of Oxford, United Kingdom

Anna Kijas
kijas@bc.edu
Boston College, United States of America

We use the phrase "high-resolution musicology" to We use the phrase "high-resolution musicology" to refer to the increasing level of detail being captured, processed, and displayed in digital tools for cataloguing and searching early music sources. These tools are made possible by the widespread availability of several technologies. High-quality digital imaging provides extremely detailed views of source materials. Co-operative cataloguing efforts have captured page-level information, allowing work-level data to be linked from one source to another in relational databases. The Music Encoding Initiative is driving renewed efforts to encode the underlying semantics and process of the music notation rather than treating notation simply as a visual medium. The papers presented in this session will demonstrate the convergence of digital imaging, data modelling, detailed cataloguing, and semantic notation encoding in medieval musicology, and will relate experiences and lessons learned in building these tools.

In *The Burns Antiphoner – From manuscriptto interactive resource,* Ben Florin and Anna Kijas report on a new project based around a 14th-century source, the Burns Antiphoner. In this project they focus on integrating and delivering images, metadata, and recorded audio and video to provide an in-depth look at the contents and experience of this particular source. Using high-resolution web-based image display technologies, they provide users with the ability to view small details on the page images, while integration with data collected in the CANTUS database provides a detailed view of the musical contents of each image, including liturgical function. A novel in-browser search system based on Lunr.js provides users with full-text and field-based search capabilities. Finally, Florin and Kijas explore the possibility of using Music Encoding Initiative (MEI) standards to encode and render the musical incipits, providing a valuable and open data source for publication and re-use. Karen Desmond further explores the use of MEI for medieval musicology in *Measuring Polyphony: A project to encode the semantics of the context-based (and under-prescriptive) notation of late medieval music.*

Previous projects have involved an implicit or explicit translation to either a purely visual medium (for the purposes of print or display), or to a system of notation encoding that treats medieval music as though it were written using modern notation. This paper explores how medieval music might be semantically structured using the Music Encoding Initiative, such that figures representing time and proportion in the music remain contextually related. This work represents a new direction in historical music notation encoding, and will provide an account of the technical and representational challenges encountered in the process.

The Digital Image Archive of Medieval Music (DIAMM) is one of the longest continually-operating digital musicology resources. Recent changes to the ways in which cataloguing data is captured, presented, and made available to others, however, present new opportunities for researchers to interact with this digital collection. Andrew Hankinson and Julia Craig-McFeely describe their efforts at promoting granular data accessibility in both human and machine-readable representations, following established and emerging standards to promote data integration and re-use within other projects.

The Burns Antiphoner – From Manuscript to Interactive Resource

Ben Florin, Anna Kijas

The Burns Antiphoner (is a collaborative project between the Digital Scholarship Group and library staff at the Boston College University Libraries, musicologist Dr. Michael Noone, and several external partners, including CANTUS database staff. In the summer of 2015 we began encoding afourteenth-century Franciscan Antiphoner using the Music Encoding Initiative (MEI) standards, as well as developing an open-access site to present the manuscript as an interactive object for research and scholarly use.[1] The original Antiphoner is in manuscript form bound between leather-covered boards containing 119 parchment folios with text and notation for antiphons and responsories for the entire annual calendar of saints' days (sanctorale).

This paper will examine the workflow and development of this project and the main goals, which included:

1. Making the Antiphoner interactive by building an open access and responsive website with Diva.js, a dynamic presentation layer that can search and display content (data, notation, XML/MEI) and multimedia;
2. Using and developing open source technology to encourage and support further development in the open source community, as well as, sharing of documentation and data;
3. Contributing our data to the scholarly community through a collaboration with CANTUS, a database for Latin Ecclesiastical Chant; and
4. Developing software, workflows, and documentation that can inform future projects using similar technology.

Key infrastructure and application choices used in this project will be discussed, including the implementation of Diva.js as the presentation layer to pull in our data via a JSON file, Lunr.js to index our data directly in the browser, and the use of IIIF, which will enable interoperability of our images across different platforms and potentially allow us to test our data with optical music recognition (OMR).

The workflows for transcribing and encoding approximately 1500 incipits for our project and the CANTUS database will be explored, including some challenges that came along with this project, which was our first endeavor in using MEI. The neume notation is currently rendered as Volpiano font, but Verovio would be preferable so that the encoded incipits in MEI XML could be rendered directly. Verovio is a software that renders MEI directly in a modern browser as SVG (scalable vector graphics: XML-based vector image format). At this time, however, while Verovio can render MEI files that use mensural elements, it cannot yet render those associated with the MEI neume module. In addition, we will discuss our ongoing explorations, which include presenting the antiphoner in a multitouch environment and investigating the application of OMR on our IIIF image manifests, containing encoded text and musical incipits from the manuscript.

Measuring Polyphony: a project to encode the semantics of the context-based (and under-prescriptive) notation of late medieval music

Karen Desmond

'Measuring Polyphony' is a project that digitizes polyphonic music of the late medieval period. The goals of the project are threefold: 1) to make transcriptions and sound files of this repertory freely available online to performers, scholars, and the general public, alongside images of the original music manuscripts; 2) to encode the medieval notation in a standardised machine-readable format so that the music can be searched or analysed using current tools, and through this interoperability the data made available to other websites and applications; and 3) to streamline the processes and tools for encoding this repertory so that other stakeholders can easily and rapidly enlarge the dataset.

The availability of high-resolution images of most manuscript sources of medieval polyphony since the late 1990s, in particular through the groundbreaking open-access DIAMM initiative, and also library-based repositories such as Gallica, has heightened scholarly awareness of the importance of considering medieval music within its original material context. The major print editing projects of the mid-twentieth century divorced the musical texts from their parchment and ink origins, segregating them into composer- and genre-ordered collections, converting their notation to modern equivalents, and secreting away detailed philological considerations and any commentary pertinent to a specific manuscript or scribal practice into cryptic appendices. Now, however, the easy access to so many different music manuscripts in the (virtual) flesh has encouraged scholars to think again about the meaning of music compositions as it relates to the manuscripts in which they are found.

This project, started at McGill University, with the support of SIMSSA project and the Schulich School of Music, and now continuing at Brandeis University, leverages the potential of these rich image repositories and the availability of community-based standards for encoding music notation. While encoding music notation is considerably more complex than encoding text alone, fortunately the standards developed by the Music Encoding Initiative (MEI) community have emerged in recent years as the standard way to encode music notation documents as XML. Most of the current projects that encode music notation, however, are focused on repertories written in the neumatic notations of western plainchant, or from the common practice period of c. 1600 and later. ‘Measuring Polyphony’ is the first project that encodes medieval music as it was originally notated in mensural notation--a notation system developed in the late thirteenth century in order to more precisely denote rhythm in polyphonic music, and which continued to be used until the sixteenth century. Mensural notation presents additional difficulties for encoding since it is a context-based notation. In other words, whereas the shape of a note and its duration are in a one-to-one relation in notation from the common practice period, this is not the case with mensural notation, where the same note shape can be used in different contexts to denote different durations.

The repertory chosen for the first phase of this project is a representative sample of sixty-four motets transmitted in the major sources of French polyphony that were copied c.1300-1375. This paper will describe the process by which the mensural notation was encoded, and the encoded transcriptions (and audio realisations) displayed on the fly using Verovio, an open-source library for engraving MEI music scores into SVG.

The encodings, at present, capture the diplomatic transcription of each composition from one manuscript source: the transcriptions, however, allow for the possibility of later adding alternative readings from concordant sources (through the Critical Apparatus module of the MEI schema), and thus for the possibility of extending these encodings to produce online critical editions of these compositions. Medieval notations are often under-prescriptive, and these encodings also allow for flexibility in imposing particular pitch or rhythmic interpretations, so that future interfaces could allow users to easily switch back and forth between differing editorial interpretations. In addition, the availability of this first dataset of a medieval repertory encoded in mensural notation, along with the tools and documentation enabling others to rapidly add to this dataset, opens up new possibilities for the analysis and interpretation of fourteenth-century music, and enriching our understandings of the the visual notation systems developed to represent and record it.

Building the new DIAMM: Linking and sharing data for medieval musicology

Andrew Hankinson and Julia Craig-McFeely

The Digital Image Archive for Medieval Music (DIAMM) is perhaps the longest-running digital resource for medieval musicologists. First launched on the World Wide Web in 1998, it has provided countless researchers, performers, and enthusiasts with a glimpse into libraries’ and archives’ collections through the use of high-resolution digital imaging and authoritative cataloguing and description efforts. In November 2015, work started on the third iteration of the DIAMM web resource, with renewed efforts on providing enhanced search and discovery interfaces, data sharing, and image resource sharing, in an effort to provide a platform for new applications in digital musicology. This presentation will provide a progress report on this work, as well as introduce users to some of the new features and capabilities afforded by these developments.

Background

The DIAMM database currently catalogues 3,375 medieval and early modern manuscripts of monophonic and polyphonic music (excluding chant). The project has high-resolution digital images of 710 sources, from which we have permission to publish 45,626 images online. Most of these images are natural-light photographs, but for a few sources we also make available 1,733 alternate images, including those shot in alternate light sources (Ultraviolet, Infrared), with transmissive light, digitally enhanced, or digitally restored images.

In addition to descriptions and images of manuscript sources, we capture extensive data about the contents of these sources. Musical compositions and textual sources are related to the individual pages in the sources, giving us the ability to identify the presence of a particular musical work across several different sources and navigate directly to the page image where it can be seen. The roles of people and organizations may also be traced across sources. For example, a record for a given individual may link to the sources and works where he acted as copyist, owner, or composer.

With previous versions of the DIAMM website, corrections and additions to the contents were extremely slow. This was due to a number of technical challenges in the publication workflow, translating the local FileMaker database to a MySQL database for publication. In the new version, we have moved to an online updating model, where content updates may be proposed by our community of users, after which they are reviewed for inclusion and made available immediately, with appropriate credit for the change acknowledged on the object record.

Addressability and Linked Data

The central organizing principle behind the new DIAMM website is the creation of unique, understandable, and resolvable resource URLs following the principles of Representational State Transfer (REST, see Fielding, 2001). A resource is defined as the objects, or "things" in the database: A manuscript source, a person, a geographic area, and so on. The URLs are designed to function as permanent and globally unique identifiers for both human- and machine-readable representations of these resources. To give an example, the URL https://www.diamm.ac.uk/sources/202/ refers to the record for the Eton Choirbook (GB-WRec MS 178). By default, loading this URL in a web browser will return the HTML version of the record for human readability. However, a machine-readable representation of this data is also available at the same URL using "content negotiation," a process through which different representations of the same record may be requested when the URL is retrieved. Asking for the content type of "application/json" will return the machine-readable JavaScript Object Notation (JSON) representation of the same data.

We designed this URL scheme to promote integration of DIAMM data within third-party applications. In many cases, DIAMM records are the only digital representations of a manuscript, person, or organization. A canonical and globally-unique URL record allows these entities to be represented in other applications, and the data we have captured about these objects can be integrated within other applications, enhancing and extending the qualities and relationships of these objects.

The International Image Interoperability Framework

The International Image Interoperability Framework (IIIF, Snydman et al, 2015) is an emerging standard, promoting a common format for delivering and structuring digital image collections. The IIIF specification is composed of two primary recommendations: The Image API, and the Presentation API. The Image API defines a common URL structure for addressing images, or regions of images, and delivering these images to a user's browser. The Presentation API defines a data structure for combining sequences of images into a coherent collection. For example, the images representing the pages of a book would be contained in a Presentation API document. The Presentation API also contains further functionality for delivering document metadata, such as bibliographic information or document structure (e.g., the pages representing the start of chapters in a book or sections in a newspaper).

For the DIAMM website we provide IIIF Presentation API documents for all sources for which we have images, and deliver our images through a IIIF Image API-compatible image server, the IIP Image Server (Pillay, 2016). Due to licensing restrictions with our partner libraries, however, the availability of images and IIIF manifests are restricted to our users who have signed in. In the next phase of our project, we hope to review these agreements to facilitate open delivery of source images to the wider public.

In addition, we have integrated IIIF manifest viewing into the source records we publish. With this change, we are able to incorporate images from other libraries' collections into our own without needing to host and re-deliver these images from our own website. If, for example, a library publishes a IIIF manifest for a particular source, we can display these images immediately on our website, with the image content delivered from the library directly. We hope this will greatly reduce the latency with which we can make images available to our community of users, and we encourage all libraries and archives to ensure their image collections are made available as IIIF-compatible sources.

A new platform for medieval musicology

The new version of the DIAMM website will continue to serve as an important destination site, available for consultation and discovery for our community of users. With the enhancements described in this presentation, however, it will also begin to serve a new and more active role within the field of digital musicology. Using DIAMM data, and providing machine-readable and addressable resources, the data from the DIAMM site can have a life outside of our own website. We anticipate that new capabilities in data availability, and a standards-based approach to content addressability and machine readability will promote re-use and integration ("mash-ups") within other tools.

Bibliography

Fielding, R. (2001). "Architectural Styles and the Design of Network-Based Software Architectures," (PhD diss., University of

California, Irvine)

Franciscan Antiphoner (sanctorale). John J. Burns Library, Boston College, University Libraries. http://hdl.handle.net/2345/2231.

Pillay, R. (2016). IIP Image Server. http://iipimage.sourceforge.net/ (accessed 31 October 2016).

Snydman, S., Sanderson, R., and Cramer, T. (2015). The International Image Interoperability Framework (IIIF): A Community & Technology Approach for Web-Based Images. Presentation given at the Archiving Conference, Los Angeles, CA. 19–22 May.

Research Computing's Demand for Humanists, and Vice Versa

Quinn Dombrowski
quinnd@berkeley.edu
UC Berkeley, United States of America

Tassie Gniady
ctgniady@iu.edu
Indiana University, United States of America

Megan Meredith-Lobay
megan.lobay@ubc.ca
University of British Columbia, Canada

Jeffrey Tharsen
tharsen@uchicago.edu
University of Chicago, United States of America

Lee Zickel
lxz11@case.edu
Case Western Reserve University
United States of America

Over the last five years, research computing has undergone a shift from focusing on running infrastructure for highly technical researchers, primarily in the sciences, to supporting medium- to large-scale computational needs across a wide range of disciplines, where practitioners fall across the spectrum of technical proficiency.

This shift in approach has opened up new opportunities, both for scholars whose research questions are facilitated by computationally intensive algorithms (e.g. photogrammetry, natural language processing, OCR at scale), and for humanists in support positions that require translation between technical and non-technical audiences. This panel will discuss opportunities for research and career development through the perspectives of research IT staff whose backgrounds in both humanistic inquiry and computational and digital methodologies enable them to engage in outreach to, training of, and support for humanities researchers. It will also provide practical suggestions for humanists who could benefit from research computing resources but find it difficult to navigate the expectations of research IT organizations.

The communication, writing, and teaching skills cultivated through an advanced degree in the humanities align with employment trends in research IT groups. While some familiarity and comfort with computation is generally a requirement for working in research IT, advanced programming or system administration skills are crucial for a minority of positions in these groups. As the mandate of research computing groups has expanded, it has become increasingly clear that successful research computing programs require documentation that is comprehensible to a non-technical audience, hands-on workshops for non-specialists, and staff capable of understanding the fundamentals of researchers' work and identifying effective approaches to meeting their computation needs. To that end, in 2014 the US National Science Foundation funded the Advanced CyberInfrastructure - Research and Education Facilitators (ACI-REF) program, which developed a cohort of computation "facilitators" across a number of universities who could "maximize the impact of investment in research computing" by "assist[ing] researchers in leveraging ACI resources for themselves" and sharing solutions among participating institutions (NSF 2014). This approach to providing support has been highly influential among campus research IT organizations, and related workshops such as the ACI-REF virtual residency (featuring plenaries such as "Effective Communication: How to talk to researchers about their research" and "Writing Grant Proposals") have drawn large crowds (Neeman et al. 2016).

Similarly, the Extreme Science and Engineering Discovery Environment (XSEDE), funded by the National Science Foundation, has begun specifically reaching out to disciplines that have been typically under-represented in the high performance computing arena, including those in the humanities. XSEDE has a program called Extended Collaborative Support Services (ECSS) that partners humanities scholars and others from under-represented areas with computing professionals who can help them achieve their desired research objectives (Wilkins-Diehr et al. 2016). This has been both necessary and fruitful particularly in areas that are focused on data analytics (such as video analysis, image analysis, network analysis, etc.), as opposed to the more traditional simulation and modeling done by scientists and engineers. XSEDE's Novel and Innovative Projects group includes a specialist in digital humanities, as well as specialists in pertinent related areas such as big data analytics.

Along the same lines, in January 2014, Compute Canada hired a full-time Digital Humanities coordinator to lead a national team supporting researchers wanting to engage with national advanced research computing (ARC) resources. Compute Canada, a national non-profit organization incorporated in 2012, plans and oversees pan-Canadian ARC resources used for big data analysis, visualization, data storage, software, portals and platforms for research computing serving Canadian academic and research institutes. The national support team for Compute Canada now consists of the full-time DH coordinator, and a large geographically dispersed team that meets bi-weekly to coordinate national initiatives like training at the Digital Humanities Summer Institute and national competitions such as the recent partnership between Compute Canada and the Canadian Social Sciences and Humanities Research Council. The DH team also meets locally with humanities researchers at their own institutions to help them leverage national infrastructure, and shares experiences and advice back to the national team to help in developing tools, services, and training opportunities that will benefit the national DH community (Simpson 2015).

One of the interesting ramifications of these changes in IT staffing trends is the addition of humanists to IT-based groups in high performance computing, cyberinfrastructure, visualization, and data architectures--humanists who are called upon to maintain both a deep understanding of computational systems, as well as track the needs of scholars in the humanities, which tend to be quite different from researchers in the “hard” or social sciences. Often, they are the lone humanist in a group dominated by engineers and computer scientists. There is a clear need and desire to expand the use of computational resources into "non-traditional" (i.e., non-hard-sciences) disciplines, both to justify an institution-wide investment in research computing, and as a way of building institutional capacity for supporting digital humanities scholarship (as described in the forthcoming 2017 ECAR/CNI white paper on institutional digital humanities support). Nonetheless, institutions are struggling with translating their services into comprehensible and relevant offerings for humanities researchers. Finding effective means of supporting these researchers within the traditional model of the research IT group has also been a challenge, given the ways that it differs from library-based support models with which scholars are more familiar. Panelists will reflect on projects they have worked on that successfully bridged the humanities-research computing divide, to the benefit of both groups.

As one example, at Indiana University, a workflow for teaching text analysis with R has been developed that uses web-based Shiny scripts to introduce an algorithm, highly annotated RNotebooks explaining each line of code, lightly annotated RScripts allowing for remixing and adaptation, and, finally, RScripts to leverage multicore environments. The scripts use data both from literature and Twitter, and these tutorials consistently draw the highest attendance in the series DH for Humanists, held throughout the semester. Individual research projects using more sophisticated algorithms such as NER and LDA have also grown out of this project.

As another example, at the University of Chicago, the recently-developed Visual Text Explorer provides a new type of framework for reading texts along with a range of user-customizable analytics, allowing for simultaneous close and distant reading. Other humanities research computing projects include web-based data-driven animated interactive mapping systems, tools for comparative sequence analysis across literary corpora, and automated aggregators of ranges of specific secondary data sources to inform the reading of specific texts/types of texts, all of which require no knowledge of computer programming by the user but nonetheless leverage research computing resources.

At UC Berkeley, 3D modeling work by Near Eastern Studies scholars that uses photogrammetry software running on the high-performance compute cluster has been helpful for testing new Graphics Processing Unit (GPU) nodes in the cluster.

Finally, panelists will provide recommendations for how humanities scholars can translate their research projects in ways that will make them more comprehensible and compelling for institutional research IT groups that may not have a humanist on their support staff.

Panel participants

Quinn Dombrowski is the Digital Humanities Coordinator in Research IT at UC Berkeley. Research IT supports research data management, museum informatics, and computationally intensive research across all domains. She is the author of *Drupal for Humanists* and has an MA in Slavic linguistics and an MLIS from the University of Illinois.

Tassie Gniady is the manager of the Cyberinfrastructure for Digital Humanities Group at Indiana University. The CyberDH group focuses on workflows for text analysis and photogrammetry. She also teaches an Introduction to Digital Humanities course in the Information and Library Science School. Tassie has a Ph.D in early modern literature from UC-Santa Barbara and an MIS from Indiana University.

Megan Meredith–Lobay is the Digital Humanities and Social Sciences Scientific Analyst for the University of British Columbia’s Advanced Research Computing Department. She is also part of WestGrid and Compute Canada, the Canadian HPC national infrastructure platform. Megan has a PhD from the University of Cambridge Department of Archaeology in which she explored the early Christian archaeology of Argyll, Scotland using GIS and early online archaeological databases.

Lisa M. Snyder is the director of Campus Research Initiatives for UCLA’s Office of Information Technology, and manager of the GIS, Visualization, and 3D Modeling group for the Institute for Digital Research and Education. She has a Ph.D. in Architecture and teaches Virtual Reality and 3D Modeling in UCLA’s digital humanities program.

Jeffrey Tharsen is Computational Scientist for the Digital Humanities at the University of Chicago where he is the lead technical domain expert for digital and computational approaches to humanistic inquiry. Jeffrey has a Ph.D. from the University of Chicago's East Asian Languages & Civilizations department, specializing in the fields of premodern Chinese philology, phonology, poetics and paleography.

Bibliography

National Science Foundation. (2014) “Advanced Cyberinfrastructre - Research and Educational Facilitation: Campus-Based Computational Research Support.” https://www.nsf.gov/awardsearch/showAward?AWD_ID=1341935

Neeman, H., et al. (2016) "The Advanced Cyberinfrastructure Research and Education Facilitators Virtual Residency: Toward a National Cyberinfrastructure Workforce". In *XSEDE16 Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*. ACM: New York. DOI: 10.1145/2949550.2949584

Simpson, J. (2015). “Building Support for Digital Humanities”. *Compute Canada* blog. https://www.computecanada.ca/blog/building-support-for-digital-humanities/

Wilkins-Diehr, N., et al. (2016)“An Overview of the XSEDE Extended Collaborative Support Program” in *High Performance Computer Applications*, ed. Isidoro Gitler and Jaime Klapp. Springer Link: 2016. http://link.springer.com/chapter/10.1007%2F978-3-319-32243-8_1

Alternate Histories of the Digital Humanities

Amy Earhart
aearhart@tamu.edu
Texas A&M University

Steven Jones
stevenjones@usf.edu
University of South Florida

Tara McPherson
tmcphers@usc.edu
University of Southern California

Padmini Ray Murray
p.raymurray@gmail.com
Srishti School of Art, Design and Technology

Roger Whitson
roger.whitson@wsu.edu
Washington State University

Recent work in the digital humanities has moved away describing the digital humanities as a “big tent,” to quote William Pannapacker’s famous 2011 post. Taking inspiration instead from the multiple histories and temporalities of media archaeology, such research emphasizes the local contexts where technological and institutional history take place. Matthew Kirschenbaum’s identification of the digital humanities in 2014 as a “discursive construction” that ignores the “actually existing projects” of the field set the stage for scholars to rethink how the digital humanities conceptualizes its work and its history (“What Is” 48). More recently, in the introduction to *Debates in the Digital Humanities 2016*, Matthew Gold and Lauren Klein use the scholarship of Rosalind Krauss who, in 1979, described art history as emerging as “only one term on the periphery of a field in which there are other, differently structured possibilities.” Whereas Krauss saw this as a failure of art history, Gold and Klein celebrate the multiplicity of what Patrik Svensson calls a digital humanities that is less a tent and more a disaggregated “trading zone” of various interests and disciplines. Instead of a transcendent, disciplinary category, the digital humanities emerges as an imminent set of assemblages and rhizomatic localities — converging in some places, diverging in others.

This panel of short papers intervenes in the discussion of an imminent digital humanities by describing several actual alternate histories of the field. All of the thinkers for this proposed panel have sketched variations on digital humanities history in the past. Steven Jones begins his book on Roberto Busa, for example, with an extended discussion of the “multiple potential continuities” existing beside the mythological figure as providing a possibility for “better historical understanding” (16). While Amy Earhart’s work historicizes digital literary studies in America through the work of the new historicism, Tara McPherson sees it in the screen cultures of media studies, Roger Whitson points to the publics outside academia invested in steampunk and other nineteenth-century sources, and Padmini Ray Murray explores the repurposing practice of jugaad in India. Such alternate histories point not to a denigration of the meaning of the digital humanities as a disciplinary field, but rather describe — as Lori Emerson says about media archaeology — each strand as “one possibility generated out of a heterogeneous past.” Each of the presenters will spend 10 minutes discussing how DH can be historicized using various disciplinary, national, and outer-institutional contexts.

Activism in Digital Humanities: Complicating Community, Technology, and Open Access

Amy Earhart

Much of our history in digital humanities has focused on proving that our work has legitimacy within the academy. As I have argued in other publications, the digital humanities has been critiqued as a regresive field, particularly in terms of its approach to cultural studies, and, at the same time, as a challenge to traditional humanities (“Futures”). Key to this simplistic critique of digital humanities is a representation of the digital humanities as a monolithic structure. As part of a panel which reveals the multiple histories of digital humanities, this paper will chart the alternative history of activism and community/academic partnerships in the digital humanities.

Arguing that critiques of digital humanities are ahistorical, the paper will focus on the connection between activism and community in the early digital humanities. For example, the public/academic focus of early digital humanities work has direct ties to what we now call public digital history. Douglas Seefeldt and William G. Thomas have argued that the future of digital history "invites students and the public into the digital process," yet this is actually not a future goal. It is our past and connects to a long historical interest in digital humanities as activism and a means of creating community partnerships.

Of particular focus, in the paper, are projects that bring scholars inside the academy into partnerships with community groups, such as the early *NativeWeb* or *eBlackStudies*. While such early projects are often viewed as retrograde technologically and often dismissed from our dh genealogy, they offer an alternative history of the way that technologies are used in service of particular fields within the academy. At the same time, such projects are interested in bridging the divide between the academy and the community and serve particular activist agendas. While there are some forms of digital humanities that reject a focus on cultural studies, this branch of digital humanities centers political activism and critiques of race, class, sexuality, and gender within its approach.

The paper will also focus on the way that technology is imagined in the various lineages of digital humanities. In the line of activist projects that the paper examines, technologies are decentralized, often out of the box, and less interested in innovation than in, say, current large corpora data mining projects. Too often "simple" technological projects are dismissed as not digital humanities, even when the theoretical usages of technology in relationship to humanities questions are innovative and forward thinking. Instead of accepting techno progressivism, scholars in digital humanities need to apply the full spectrum of humanities critique to the treatment and use of technology. For example, scholars have a responsibility to address the ways that technological specifications might force western representations of knowledge onto materials of cultural expression that do not use such systems. Projects such as the *Tibetan and Himalayan Library's (THL)* use of TEI/XML provides one example of how we might proceed. The *THL* has considered how the understanding of time might be culturally constructed and, as such, has revised the TEI/XML coding to reflect time from the perspective of the Tibetan culture rather than imposing western understandings of time through technological standards.

The history of activist digital humanities projects reminds us to think about how the exploitation of data is related to historical exploitations of people(s), to reconnect the digital with embodied experience. Mark Turin notes, "archives become more complex when the 'documents' in question are representations of human 'subjects,' as was the case for the ethnographic archives in which we were interested, including photographs, films, sound recordings and field notes on people's lives, their cultures and their practices" (453). Documents are never devoid of embodiment, as we might never use the term exploitation of data without understanding that, eventually, exploitation of data has real impact on individuals and communities. A division of human subjects and documents leads to problematic interactions with those who we are working to digitize. We need to think about how our data embodies experience.

The paper will close by focusing on the way by which ideas of open access are culturally constructed. Activist projects complicate the adage "information wants to be free," reminding digital humanities practitioners that the model of broad 'access' that often motivates western digitization efforts does not apply universally." The complexities of technology as represented by such practitioners are central to digital humanities.

Roberto Busa, S.J., and Humanities Computing: Complicating the Origin Story

Steven Jones

The Jesuit scholar, Roberto Busa, is often called the founder of humanities computing. In fact, starting as early as 1949, he collaborated with IBM to perform experiments using suites of punched-card machines. These punched-card data systems—with their plug-board setups, clacking machinery, and flurries of perforated rectangular cards—were developed for business accounting and tabulating, and adapted for government censuses, defense calculations, archival management, and information processing of all kinds. These systems coexisted for many years with electromechanical calculators and electronic computers, helping to define, delimit, and shape the possibilities for research applications, including humanities research applications like Father Busa's. Because the card systems were eventually connected to electronic computers, they've become part of the story of humanities computing. But in many ways, the first decade of humanities *computing* can more accurately be described as an era of humanities *data processing*—in the historically specific and contextually rich sense of the term.

My historical work on Roberto Busa's data processing has drawn on a key premise of media archaeology: that technology doesn't "evolve," or "descend," in a linear way. As Michel Foucault asserted, genealogy (in the sense he used the term) cannot be figured by strictly logical trees of descent, as in the "evolution of a species." It's a way of viewing events in their "proper dispersion," including the "minute deviations . . . complete reversals—the errors, the false appraisals, and the faulty calculations . . . the exteriority of accidents" that constitute history. When it comes to Busa, so often treated as the "founding father" of humanities computing and digital humanities, one way to complicate the origin myth is to pay attention to the "exteriority of accidents" that shaped the received story, to tell the story slant, as it were, by looking at what Steven Johnson has called the "adjacent possibilities," even the dead ends or paths not taken that nevertheless help us to understand what was done.

One example is the fascinating Microfilm Rapid Selector machine, which Busa briefly considered but from which he swerved away. Based on an experimental design by Vannevar Bush for the famous memex, the prototype was viewed by Busa in operation at the library of the Department of Agriculture in 1949. It offered a competing paradigm (both technically and institutionally) for information processing and retrieval. Or, the large-scale photo-mainframe IBM SSEC, which Busa saw working at IBM in 1949-1952 but was unable to use himself (since it was dedicated to scientific and industrial applications). It nonetheless inspired his thinking about the nature and scale of his linguistic data. Its existence as a kind of adjacent possibility is a useful reminder both of the institutionalization of the "two cultures" of science and the humanities at mid century, and, at the same time, of the artificiality of the categories. Or, take the Dead Sea Scrolls project, which Busa's "lab" undertook but could not complete (and the remnants of which remain in the Busa Archive in Milan), but which revealed some of the limits of the idealized "computerized philology" that Busa was pursuing at the time.

My work on Father Busa, IBM, punched-card machines, and large-scale calculators has been inspired by Geoffrey Rockwell and Stéfan Sinclair, who have called for a media-archaeology approach to the technologies of the mainframe era. I've also drawn on a related approach, platform studies, which looks at individual platforms in their multilayered material particulars. Because media archaeology looks at forgotten or discounted technologies (presumed to be superseded by what has come to dominate the present), and replaces a triumphal narrative of technological progress with messier stories, it can usefully check and complement the laser focus of platform studies. Together, they allow for richer, more detailed views of the changing cultural and historical conditions within which technologies emerge and jostle for prominence.

Theory/Practice: Lessons Learned from Feminist Film Studies

Tara McPherson

This talk investigates possible relationships of theory to practice within digital humanities and media studies and also calls for a politically engaged approach to both fields. It seeks to move beyond the binary framing of the DH slogan, "less yack, more hack," by arguing for a more integrative and dialectical melding of making and critique. In particular, I turn to feminist film studies of the late 1970s to examine an earlier moment in media studies that sought to integrate media production, distribution, theory, and pedagogy toward expressly political ends. In conversation with contemporary feminist scholarship in new materialisms and digital media studies, I argue that practices of making can and should enrich our theoretical and discursive endeavors within feminism.

In a recent article taking up the phrase "less yack, more hack," Claire Warwick helpfully suggests that increased focus might be paid to the qualifiers "less" and "more" rather than to a binary opposition between "yacking" and "hacking" (538). I agree. We can then see yacking and hacking as held within a productive and dialectical relation. To take this line of thinking further, we might not even focus on "less" or "more," as if the relationship between theory and practice can be reduced to balancing a formula. Instead, we might understand the two terms to be tied together in a productive and iterative friction. The tensions between "yack" and "hack" are not, perhaps, all that unique to the digital humanities. They exist across the university in structures that make it hard to combine theory and practice in our curricula, evaluation and promotion structures, disciplinary methodologies, and privileged forms of scholarly output.

As digital humanities scholars have struggled with the right balance of yack and hack, broader debates have emerged about the relationship of theory to practice across the academy. If these tensions have simmered just below the surface of disciplines for much of the twentieth century, then the digital turn has reanimated such debates in new ways in the new millennium. Returning to earlier moments of practice within and beyond the academy can provide valuable lessons for DH today.

The wedding of theory and practice was crucial to the formation of feminist film studies as a field over forty years ago. We see this joining quite literally in the title of key essays such as Claire Johnston's "The Subject of Feminist Film Theory/Practice." The "/" signals a hybrid practice beyond the "and." Published in the journal *Screen* in 1980, the piece reports on the Feminism and Cinema Event held at the 1979 Edinburgh Film Festival. Johnston writes that, "throughout the week, emphasis was placed on the need to locate feminist politics within a conception of film as a social practice, on the dialectic of making and viewing and on film as a process rather than an object" (27). The early history of feminist film theory models a vibrant relationship of making to critique. Many feminist film scholars engaged in a form of inquiry that understood theory and practice to be constitutive of one another. These insights were often born of a deep exploration of the material forms of cinema—explorations enriched by practice. Such work is vitally important terrain for digital humanities, leading us to ask how our machines encode culture in very particular and often damaging ways while also perhaps signaling an enhanced role for artists and designers within DH endeavors.

Decolonising Design: Critical Making and Jugaad in India

Padmini Ray Murray

This talk will cover the establishment of the first Masters-level digital humanities programme in India, as well as a digital humanities research agenda that contextualizes and embeds such work in an Indian environment. Ray Murray will discuss the shift that this unique context generates, in terms of necessarily moving away from modeling such work on paradigms established in Anglo-American institutions, and her work towards the creation of a locally reflexive practice which responds more appropriately to its conditions. Drawing on the work of architectural historian Arindam Dutta, Ray Murray will historicise these arguments, demonstrating how design as a discipline is implicated in the work of colonialism, and how the praxis of critical making (as formulated by thinkers such as Matt Ratto and Garnet Hertz) can contribute both to decolonising design and more broadly, humanities scholarship in India, as well challenging traditional institutional frameworks that are the legacy of colonial education. Ray Murray will demonstrate how critical making is a particularly useful mode of inquiry in a context where digital humanities work is relatively nascent, in order to supplement and inform an emergent narrative of the history of what might be considered digital humanities in India.

Radically different technological and infrastructural conditions as well as historical mean that this narrative diverges from those which underpin established histories of Anglo-American digital humanities, and Ray Murray will explicate on how this difference necessitates alternative methodological approaches in order to reconstruct alternative histories. The work of Jentery Sayers and others at the Maker Lab at the University of Victoria on their cultural kits for history, while emphatically not exercises in replication and more in remediation, and in foregrounding "how the past is interpreted through present conditions, exhibiting history as a collection of refreshed traces, with both loss and gain" often relies on historical material culture such as patent documents, illustrations, artefacts in order to inform their creation, much of which is conserved and made available by Victorian values of empire-building and taxonomic collection. In contrast, the history of indigenous technologies in India is patchy and often obscured by more visible archival material which asserted colonial structures of oppression, complicating the use of a mode of inquiry such as Sayers et al's cultural kits for digital humanities work in India.

In addition, in a country like India, where literacy is still at a premium, design and making privileges the value of other forms of knowledge found in communities, such as crafts or indigenous traditions. Ray Murray will thus demonstrate to how conceiving of critical making in the tradition of practices such as *jugaad*, an indigenous combination of making-do, hacking, and frugal engineering makes for a contextually appropriate intervention in understandings of the digital humanities, and allows for a more politically nuanced view of tools, materials and the conditions of production that have laid the foundations for digital humanities scholarship going forward. In closing, Ray Murray will discuss how contemporary design education privileges "solution-ing", which anticipates a model of consumption rather than co-creation, tracing a trajectory from colonial ambition to neoliberal inevitability—and how digital humanities thinking discourages this mode by with its legacy of interpretation that discourages a one-size-fits-all response. Ray Murray will conclude by asserting the uses of creating a useful methodology to turn the lens of scrutiny upon digital artefacts and activities as well as being observant of different materialities and modalities of knowledge production in order to both historicise and limit the over determined nature of the digital.

Alternate Histories: Steampunk Fandoms and Digital Humanities Publics

Roger Whitson

The digital humanities is often characterized as dedicated to making scholarship publicly accessible. Yet accessibility is only one way to pursue a public digital humanities agenda. Another method leverages the complicated history described by media archaeology to highlight how various publics outside of University settings are already constructing digital humanities projects of their own. Jussi Parikka begins *What is Media Archaeology?* with an extended consideration of steampunk as an exemplary media archaeological practice, arguing that it falls outside of mainstream digital methodologies and is what Deleuze and Guattari call a "nomadic, minor science": a set of quirky hacker techniques whose innovations are appropriated by the more economic powers of the state (qtd. in Parikka 168). As with any manifestation of what Deleuze and Guattari call "royal science," or a hegemonic system relying upon the appropriation of nomadic practices, steampunk creates a tension between such minor sciences and their corporate and academic use. For every fascinating gadget produced by steampunk fans, there are also corporate phenomena like Justin Bieber videos featuring joyless representations of steampunk automatons whose cogs are appropriated only to sell more albums.

This talk explores a set of steampunk projects from fans in order to show how their methodologies constitute an alternate history of the digital humanities rooted in the practice of public hobbyism. One example of this steampunk hobbyist practice is Tim Robinson's 2007 build of a Babbage's *Difference Engine No. 1* from parts manufactured by the toy company Meccano. Robinson says that he was intrigued by the brand's claim to "do something real," and the tactile quality of Meccano parts mediates this sense of reality: the cold metal, the round rivets, the clicking of metal rods as they are moved by other parts. The machine's design is based upon Babbage's first engine and calculates numbers up to four digits and three orders of difference. It is composed of several ratchet wheels, each with 20 teeth and which are covered by printed tape showing numbers from 0 to 9. While visiting Robinson's website, you can find descriptions of his nostalgia for the toy company, which he describes as helping him build "the machines of my youth" — including "astronomical clocks, orreries, looms and other textile machinery [...] and perhaps most enduring, the differential analyzer (and analog computer)."

Robinson's project exists within a wide variety of other steampunk gadgets that express both nostalgia for various parts and fascination with methods of building: from other models of the difference engine, like Andrew Caroll's version created with Lego parts and rubber bands; to the varied projects of *The Steampunk Workshop*'s Jake von Slatt — who rescues available parts from junk yards and repurposes them into workable Steampunk RVs (Recreational Vechicles), Wimshurst Influence Engines, and even a Stroh violin with an amplifying horn and aluminum diaphragm. For me, such projects underscore Matthew Kirschenbaum's argument that hobbyist activities enable the digital humanities to value "the unapologetically small, the uncompromisingly local and particular" ("Ancient" 196). Yet, steampunk hobbyism also enables a different understanding of the role various publics who engage in such activity play in the digital humanities as a field.

Many digital humanities projects envision the public as a homogeneous entity who acts primarily as an audience or — in some cases — a collaborator for what ends up being essentially a scholarly act. The sheer diversity of steampunk fandom, on the other hand, resists such an easy or homogeneous definition. While some aspects of steampunk fandom act, as China Mieville has observed, as forms of nostalgic imperialism; or as Charles Stross claims, as romances with totalitarianism, other fans use steampunk to imagine histories where the Industrial Revolution happened in Africa or China rather than in Europe. Miriam Rocek dresses up as a time-traveling "Steampunk Emma Goldman" and participates in protests like Occupy Wall Street. Lisa Hager, meanwhile, uses her steampunk persona to advocate for gender neutral bathrooms. Such diversity underlines the need to understand how steampunk and the digital humanities communities exist as discrete assemblages, rooted in the politics of the communities practicing them. While this talk will cover mainly hobbyist projects within steampunk fandom, it will contextualize that work with a multiplicity of various local practices. All of these practices, I argue, extend to the digital humanities as a field — which is less a big tent and more a massive assemblage of becoming, branching, and multiplicity.

Bibliography

Dutta, A. (2009). "Design: On the Global (R)Uses of a Word." *Design and Culture*, vol. 1, no. 2, pp. 163–186.

Earhart, A. (2016). "Digital Humanities Futures: Conflict, Power, and Public Knowledge." *Digital Studies/Le Champ Numerique*, 2016.

Earhart, A. (2015). *Traces of the Old, Uses of the New: the Emergence of Digital Literary Studies.* University of Michigan Press, Ann Arbor..

Emerson, L. (2014). *Reading Writing Interfaces: from the Digital to the Bookbound*. Minneapolis, University of Minnesota Press

Gold, M. K., and Klein, L.F. (2016) *Debates in the Digital Humanities 2016*. Minneapolis, University of Minnesota Press

Hertz, G. (2012). "Introduction: Making Critical Making." *Critical Making*, Telharmonium Press, Hollywood, 2012

Johnson, S. (2014) *How We Got to Now: Six Innovations That Made the Modern World*. New York, Riverhead Books

Johnston, C. (1980). "The Subject of Feminist Film Theory/Practice." *Screen*, vol. 21, no. 2,pp. 27–34

Jones, S.E. (2016). *Roberto Busa, S.J., and the Emergence of Humanities Computing*. London, Routledge

Kirschenbaum, M. (2016) "Ancient Evenings: Retrocomputing in the Digital Humanities." *A New Companion to Digital Humanities*, Edited by Susan Schreibman et al., Wiley Blackwell, London, pp. 185–198

Kirschenbaum, M. (2014)"What Is 'Digital Humanities,' and Why Are They Saying Such Terrible Things About It?" *Differences: A Journal of Feminist Cultural Studies*, vol. 25, no. 1, pp. 46–63

McPherson, T. (2017). *Feminist in a Software Lab: Difference Design*. Harvard UP, Cambridge, 2017

Pannapacker, W. (2011). "'Big Tent Digital Humanities': A View from the Edge." *The Chronicle of Higher Education* , 31 July 2011

Parikka, J. (2012) *What Is Media Archaeology?* Cambridge, Polity Press, 2012

Patrik, L.E. (2007) "Encoding for Endangered Tibetan Texts." *Digital Humanities Quarterly*, vol. 1, no. 1, 2007

Ratto, M. (2011) "Critical Making." *Open Design Now: Why Design Cannot Remain Exclusive*, Edited by Bas Van Abel et al., Bis Publishers, Amsterdam, 2011, pp. 202–213.

Ray Murray, P., and Hand, C. (2015) "Making Culture: Locating the Digital Humanities in India." *Visible Language*, vol. 49, no. 3, 0ADAD, pp. 140–155.

Robinson, T. (2008). "Robinson's Difference Engine No. 1." *Robinson's Meccano Building Site*, 24 Feb

Rockwell, G. and Sinclair, S (2014). "Past Analytical: Towards an Archaeology of Text Analysis Tools." *Digital Humanities 2014 Conference*, 2014.

Sayers, J. (2015) "Kits for Cultural History." *Hypperhiz*, no. 13.

Seefeldt, D., and Thomas, W.G. (2009) "What Is Digital History." *Perspectives on History: The Newsmagazine of the American Historical Association*, 0ADAD.

Turin, M. (2011). "Born Archival: The Ebb and Flow of Digital Documents from the Field." *History and Anthropology*, vol. 22, no. 4, pp. 445–60.

Warwick, C. (2016) "Building Theories or Theories of Building? A Tension At the Heart of Digital Humanities." *A New Companion to Digital Humanities*, Edited by Susan Schreibman, Ray Siemens and John Unsworth, Wiley Blackwell, London, UK , pp. 538–552.

Werner, S., and Kirschenbaum, M. (2014) "Digital Scholarship and Digital Studies: The State of the Discipline." *Book History*, no. 17, pp. 406–48.

Whitson, R. (2017) *Steampunk and Nineteenth-Century Digital Humanities: Literary Retrofuturisms, Media Archaeologies, Alternate Histories.* London, Routledge.

Zielinski, S. (2006). *Deep Time of the Media: toward an Archaeology of Hearing and Seeing by Technical Means*. Cambridge, MIT Press.

Imaginaires et pratiques de la culture mobile

Bertrand Gervais
gervais.bertrand@uqam.ca
Université du Québec à Montréal, Canada

Sophie Marcotte
sophie.marcottre@concordia.ca
Université Concordia, Canada

Benoit Bordeleau
litteraturelqm@gmail.com
Université du Québec à Montréal, Canada

Gina Cortopassi
cortopassi.gina@uqam.ca
Université du Québec à Montréal, Canada

Lisa Tronca
tronca-pignoli.lisa@courrier.uqam.ca
Université du Québec à Montréal, Canada

Alexandra Tremblay
tremblay.alexandra.9@courrier.uqam.ca
Université du Québec à Montréal, Canada

La culture du réseau sans fil et du partage instantané a profondément remodelé notre conception du monde, ainsi que nos pratiques d'écriture, de lecture et de spectature. L'objectif de cette table ronde est d'examiner quelques projets récents qui cherchent à exploiter et à comprendre ces nouveaux dispositifs.

La question que nous nous posons est avant tout celle des usages des technologies mobiles numériques dans le cadre des pratiques artistiques et littéraires actuelles. Comment les créatrices et les créateurs se servent-ils des médias et dispositifs numériques? De la même façon, comment nous servons-nous des dispositifs numériques comme outils de façonnement d'un discours critique sur les arts et les littératures, de recontextualisation, de mise en scène ? Comment la cohabitation entre les œuvres et les discours critiques peut-elle donner lieu au déploiement d'un imaginaire de la mobilité et de la création numérique? Il ne suffit pas qu'il y ait des technologies pour que des pratiques naissent, encore faut-il des environnements de recherche, de connaissance et de collaboration qui favorisent leur reconnaissance et leur diffusion.

Il s'agira de considérer la mobilité comme condition et facteur de création des œuvres ou comme modalité de réception. Les intervenant-e-s aborderont en ce sens les expressions de cette mobilité sous trois angles complémentaires, à savoir les dispositifs mobiles, les plateformes (Twitter, Periscope, Facebook, Facebook Live, Instagram, Snapchat, Youtube, etc.) et leurs usages et détournements.

Nous interrogerons par exemple la façon dont les plateformes mobiles se posent comme lieux d'exposition. Les internautes deviennent commissaires en ligne, agençant photographies et textes selon des critères de cohérence qui leur sont propres. Instagram ou Tumblr, en tant que plateformes centrées sur l'image, s'imposent comme des espaces parallèles à l'espace muséal où le travail d'association et de rapprochement de l'internaute se fait valoir. Les plateformes mobiles sont également le lieu d'exposition de soi par excellence. Les internautes sont invités à performer une identité projetée et régentée, encore une fois, par une ligne directrice qui révèle des appartenances et des positionnements. Les artistes de groupes marginalisés saisissent notamment cette opportunité pour bâtir des communautés en ligne liées par un discours commun. Ces considérations seront l'occasion d'explorer deux pans de la pratique qui se recoupent par un même principe, soit la mise en scène ou l'exposition en contexte numérique.

Ces questions sont étroitement liées aux enjeux politiques et identitaires en ligne, soulevés par plusieurs créateurs et créatrices. Critiques, les œuvres tendent alors à employer différentes stratégies esthétiques pour rendre compte des rapports de pouvoir, souvent invisibles, qui régissent pourtant le réseau. Nous commenterons à cet effet les œuvres qui détournent les flux du Big Data pour créer des espaces-autres, que ce soit par l'entremise du court-circuit, du piratage, de l'infiltration ou de l'appropriation.

Enfin, plusieurs thèmes et problématiques recoupent l'ensemble des présentations brièvement déclinées ci-dessus, soit l'introduction d'un nouveau rapport au temps et à l'espace, l'adaptation de nos modes de lecture et de spectature et la constitution de communautés de créateurs-rices et de chercheurs-ses.

Thèmes abordés:

1) Plateformes mobiles comme lieux d'exposition
2) Pratiques littéraires d'écriture et de lecture en contexte numérique
3) Parcours littéraires et imaginaires urbains
4) Détournements du big data
5) Représentation des communautés marginalisées en ligne

Building Space for DH Communities

Matthew K. Gold
mgold@gc.cuny.edu
The Graduate Center, CUNY, United States of America

Scot French
scot.french@ucf.edu
University of Central Florida, United States of America

Lisa Spiro
lspiro@rice.edu
Rice University, United States of America

Micki Kaufman
mickikaufman@gmail.com
The Graduate Center, CUNY, United States of America

Erin Glass
erglass@ucsd.edu
UC San Diego, United States of America

Jessica Pressman
San Diego State University, United States of America

Lisa Rhody
lrhody@gc.cuny.edu
The Graduate Center, CUNY, United States of America

Description of Session Topic

How can DH regional communities best be cultivated and sustained? This panel explores US-based digital humanities collectives that foster active communities of practice. Regional consortia are a growing phenomenon in US digital humanities, offering opportunities for loosely organized groups of DHers to share workshops, events, datasets, conferences, and news with one another. Such arrangements have the benefits of addressing institutional barriers to DH work and enabling sharing that can help address, at least in a small way, funding and infrastructural inequities that can make it hard for newcomers to begin DH work. According to John Theibault, regional consortia "can be distinguished from, on the one hand, state and national digital humanities groups that organize conferences and edit journals and require paid membership, as well as digital humanities centers located in a single institution or formally constituted groups with explicit criteria for admission, and, on the other hand, not visibly organized interactions in active digital humanities regions, even if those interactions are frequent in practice" (2016). Theibault identified 11 self-organized regional consortia not affiliated with a larger organization such as EADH, mostly in the US. Such consortia can serve an important need in connecting researchers from a range of institutions and building community more expansive than a campus and more localized than a national or international scholarly organization. As Rebecca Frost Davis and Bryan Alexander have pointed out, "large-scale multi-institutional projects aimed at building resources and pooling expertise . . . [can be] constructed to match the needs of both small liberal arts colleges and large research institutions." The building out of regional consortia, then, can help broaden the impact and audience for digital humanities work while simultaneously addressing infrastructural needs and allowing institutions to share complementary strengths.

This panel will consider how a range of regional DH groups have organized their communities, many of them through the free WordPress-based platform Commons In A Box, to build active and lasting communities of DH practice. Presentations will cover the contours and plans of specific regional communities as well as the guidance such examples may offer to academic communities just beginning to organize themselves. Attention will be paid to shared commonalities across regional organizations as well as distinctive areas of focus. Included in the panel will be a discussion of software tools that communities can use to build regional DH organizations, with a particular focus on the Commons In A Box software used by many of the groups represented on the panel. The panel will also delve into the challenges of organizing and sustaining a regional consortium, including rewarding the volunteer labor that such organizations depend upon, raising awareness of the consortium, and developing an appropriate governance model.

Though this panel focuses on US-based regional organizations, it is hoped that the panel will stir conversations on an international level about shared models of community sustenance and infrastructure. Out of this exchange, the panel hopes to build an informal network of regional DH consortia that can serve as the basis for the ongoing development of DH communities.

CBOX Development and NYCDH

Matthew K. Gold

This presentation will discuss the design, implementation, and future plans for Commons In A Box (CBOX), with a particular focus on how it is being used in the NYCDH community to foster community among New York City DH researchers and practitioners. In NYCDH, a CBOX site has been used to create groups around discrete topics such as "Digital Art History," "Librarians In DH," and the "Digital Antiquity Working Group." Each of these groups were started by members of the site and are used, along with larger public groups such as "Announcements," to foster connections across a range of NYC institutions. In recent years, the site has facilitated the planning and implementation of an annual conference that includes a week of workshops on DH subjects offered across the NYC area by a range of institutions, and an annual prize for the best graduate student DH project. The site includes a shared calendar that is populated by members and member institutions, so that it provides a quick sense of DH events in the region.

Commons In A Box is a free software project supported by a grant from the Alfred P. Sloan Foundation and maintained by the Graduate Center of the City University of New York. Commons In A Box has two elements: a plugin manager and a default theme. The plugin manager enforces version dependencies between a collection of WordPress plugins, ensuring that all such plugins work seamlessly and without conflict—a technical innovation for the WordPress plugin ecosystem. For the end-user, the experience of installing Commons In A Box is extremely smooth; upon activating CBOX, the plugin manager automatically performs what would otherwise be dozens of manual steps, including plugin installation, activation, and configuration. The CBOX theme, which is based on the design of the CUNY Academic Commons, can also be installed; when activated, it creates a Commons space that features image sliders, community elements such as lists of members and recently updated blogs in the sidebar, and group-related functionality such as "reply-by-email" group forum functionality that has been crucial to community building on the CUNY Academic Commons. The result, and the key technical innovation of the CBOX project, is the transformation of a complicated set-up and customization procedure into an easy, streamlined installation process. CBOX is currently used by a number of organizations, including the Modern Language Association, NYCDH, Texas DH, Virginia DH, and Florida DH, to build Commons spaces for DH communities.

Under the aegis of an NEH Implementation Grant, Commons In A Box - OpenLab is a new version of CBOX dedicated specifically to teaching and learning, called CBOX-OL. In this presentation, we will describe how integrating a suite of teaching-centered digital tools for content sharing and annotation into the core code of CBOX and bundling it with a core set of DH pedagogical tools ensuring that CBOX-OL communities meet accessibility standards. We will also partner with OER initiatives at partner institutions to create ways of sharing open-access humanities content within and across Commons environments that promote and reinforce sound citation practices. Finally, we have secured a commitment from Reclaim Hosting, a popular hosting service for educational institutions and digital humanities projects, to integrate CBOX-OL into its suite of easily installable software packages. Included in the new version of CBOX will be a set of teaching-related plugins as well as a suite of DH-related WordPress tools in the areas of scholarly communication, such as Braille, which translates English text into Braille; Anthologize, which enables bloggers to create eBooks from their posts; DiRT Tools, which helps community members identify the tools from the DiRT Directory they use; and PressForward, which can be used by teachers to collect content, and then to assign the selection, evaluation, and re-publication of that content with an introduction as an activity to encourage students to connect their classroom discussion with ongoing contemporary debates. All of these plugins will benefit the digital humanities humanities via their potential to enable open learning and exchange, emphasizing scholarship as an open, process-based activity, rather than a closed, solitary one.

Commons spaces don't build themselves; rather, they are cultivated and energized by participation and leadership from members. Technical development lays the foundation on which such work can happen, but the growth, evolution, and sustainability of such platforms require vision, commitment, and resourcefulness from members of the communities where they launch. This presentation will close with a consideration of the ongoing personal, institutional, individual, consortial, and infrastructural support needed to foster and maintain flourishing communities of active practitioners.

The CUNY Academic Commons

Micki Kaufman and Lisa Rhody

The CUNY Academic Commons is an academic network built at the City University of New York (CUNY) by and for faculty, graduate students, administrators, staff, postdocs, and alumni across the 24-campus system. The site serves to foster community and promote scholarship across our twenty-four campuses. With over 7,293 members and well into its seventh year, the Commons continues to evolve as a vibrant space where members connect, create, collaborate, and explore.

As the site has grown and matured, user outreach has become an increasing focus. In addition to webinars and one-to-one engagement with users and administrators of Commons groups, the team has most recently established Commons Faculty Fellowships to assist faculty in setting up Commons sites and customizing the platform to provide the right teaching tool for class needs. In conjunction with CUNY campus Teaching and Learning Centers, these outreach efforts are helping to promote the benefits of the Commons as a teaching tool across the 24 campuses.

A significant focus of the Commons development has been on personalization of the user experience. With the addition of attractive, customizable public profiles, members can now create beautiful online portfolios for their work. The high-impact header section provides an elegant profile synopsis and collapses as one scrolls down the page. Social media icons and website links help connect user profiles to a range of services on the web. Our sophisticated profile builder makes it easy to create an online CV using free-form and specialized widgets designed to highlight positions, education, publications, and interests, and the new RSS feed widget helps members include excerpts of recent posts. Likewise, the addition of Quick Links gives user profiles, blogs and groups customizable shortened links, easy to remember and ideal for business cards and CVs. Using the cuny.is/ URL shortener helps Commons users to personalize their site and to more effectively communicate their relationship to CUNY.

The Commons allows users to create and join as many groups and websites as they want. This allows users to administrate departments, teach classes, discuss and share resources on a selected topic, and connect with people sharing common interests on the Commons. Features of the Commons include the ability to have groups be public, private, or hidden and can include a host of functions including discussion forums, file storage and document collaboration, and rich email integration that allows a more email-driven, 'listserv-like' interaction with the Commons. Sites created by Commons users take many forms including personal blogs, research projects, department, class, event or conference sites, journals, reviews and news commentaries, and photo blogs, and are configurable using hundreds of themes and plugins for a wide range of visual presentations. Users also have free access to the Commons community's WordPress Help group to get help building their sites and managing accessibility.

The Commons also now includes Social Paper, a networked writing environment that enables students to compose and share all forms of their written work across classes, disciplines, semesters, and publics. Likewise, students can browse and comment on the papers of their peers. Unlike many learning management systems or course blogs, Social Paper gives students full control over the sharing settings of each individual piece of writing. Students may choose to share a paper with a professor, a class, a writing group, the public at large, or alternately, keep it private as part of their personal, in-progress, reflective writing portfolio. Additionally, while composing, students can post comments on their writing with questions mentioning other users in order to solicit peer feedback or interest. By giving students a centralized space to manage the totality of their writing, students can easily change privacy settings as they mature as writers and thinkers, develop audiences for their growing body of work, and reflectively build off prior writing.

As the Commons continues to grow, and as its development team continues to release updated versions of Commons In A Box, the site will endeavor to continue serving the needs of faculty, students, administrators and alumni across the 24 CUNY campuses -- and in the process exemplifying how an academic social network can be sustained across a university system with multiple campuses.

Texas Digital Humanities Consortium

Lisa Spiro

Even as digital humanists at Texas colleges and universities are creating significant digital humanities projects, we face common challenges, such as finding collaborators, learning new skills and developing educational programs. In the fall of 2013, Lisa Spiro (Rice), Cameron Buckner (University of Houston), and Laura Mandell (Texas A&M) discussed establishing a statewide digital humanities consortium to connect digital humanists across the state and foster collaborations. The University of Houston hosted the first Texas Digital Humanities Consortium (TxDHC) conference, co-sponsored by Rice and Texas A&M, in April of 2014. At the conference, we held an open business meeting to plan the TxDHC, which attracted 19 participants from colleges and universities across the state. We discussed common needs, including to support faculty, graduate, and undergraduate training in DH; to build community so that members are aware of projects, opportunities and potential collaborators across the state; and to gain access to infrastructure (such as Omeka) for projects (Spiro 2014). To meet these needs, we planned to develop a website, hold an annual peer-reviewed conference, and provide informal opportunities to interact, such as by publicizing visiting speakers at our home institutions. We also explored creating internship opportunities for graduate students and advocating for DH. Rather than establishing formal structures, we decided to operate as a "coalition of the willing," with decision-making by consensus. To gather additional input, TxDHC used an online survey, which had 14 respondents between April and September of 2014. Respondents ranked the following as the leading "high" priorities: foster networking (92.9%); identify researchers in Texas with common interests (71.4%); and facilitate collaborative research (64.3%).

Informed by the input received from the meeting and survey, Spiro set up a founding steering committee (SC) for TxDHC, inviting representatives from Texas universities and colleges to serve. Current SC members include Spiro (chair), Jennifer Hecker (University of Texas), Laura Mandell, Rafia Mirza (UT-Arlington), Laurel Stvan (UT-Arlington), Toniesha Taylor (Prairie View A&M), Andrew Torget (University of North Texas), and Dillon Wackerman (SMU); representatives from Southwestern University and University of Houston have also served on the committee.The Steering Committee meets via video conferencing several times each year to discuss the ongoing development of consortium and events such as conferences and webinars; we also communicate fairly regularly through email. Our mission is to " promote digital research in the humanities disciplines and facilitate interaction amongst researchers working in the digital humanities both within the state, nationally, and internationally" by connecting people, facilitating training and knowledge sharing, and raising the visibility of DH work. Membership in TxDHC is open to anyone who sets up a profile on the organization's website. Currently there are 58 "active" members and 116 members signed up for announcements.

This presentation will explore TxDHC's history, initiatives, challenges and future plans. TxDHC's core activities focus on building community across the state, including through:

- **Our website.** Texas A&M's Initiative for Digital Humanities, Media, and Culture (thanks particularly to the work of former staff member Matthew Christy) installed and hosts the TxDHC's website, using Commons in a Box to encourage collaboration. The website includes member profiles, groups focused on topics such as Training and Metadata Standards, and a calendar. We have also implemented the DiRT Tools plugin to identify tools used by TxDHC members .
- **State-wide conferences.** TxDHC has sponsored two multi-day conferences: its inaugural conference at the University of Houston in 2014 and its second conference at UT Arlington in 2015. While these conferences boasted strong programs, recruiting Texas institutions to host the event has proven challenging, especially since TxDHC lacks resources beyond endorsing the conference, publicizing it, and enlisting steering committee members to assist with it. In 2016, TxDHC shifted to a strategy of partnering with other Texas digital humanities/ digital library conferences to hold post-conference events, avoiding duplication of efforts, making it more convenient for people to attend both events, and taking advantage of cross-publicity. Recently we organized a one-day hybrid unconference/mini-conference following the Texas Conference on Digital Libraries in Austin and a THATCamp following Digital Frontiers in Houston. The mini-conference combined the best of THATCamps and more formal conferences by giving speakers fifteen to twenty minutes to set the context, then devoting the rest of the hour to discussion. While attendance at THATCamp Digital Frontiers was fairly small, the event connected participants from Rice, the University of Houston, UT Arlington, UT Austin, and local museums and explored topics such as creating an introduction to DH course and reviving DH projects.
- **Webinars:** TxDHC runs occasional web-based workshops on topics such as OpenRefine, DPLA, and using tools like Omeka or Hypothes.is in the classroom. To encourage interaction and build community, we employ a video conferencing platform so that participants can see each other's faces, and we try to set aside time at the end of each event for information sharing.

TxDHC faces several challenges:

- **Raising awareness of its existence.** At the recent Digital Frontiers conference, it became clear that many attendees didn't know about TxDHC. The organization is primarily publicized through its events, website, word of mouth, and Twitter, but more outreach to DH groups across the state is needed.
- **Accomplishing its vision with few resources.** Lacking a budget or staff, TxDHC depends on the time and commitment of its hard-working volunteers, particularly its steering committee members. Of course, these volunteers face competing demands on their time, so consortium work is fit in as it can be.
- **Keeping the website updated and encouraging people to make full use of it.** Ideally, members of TxDHC would add events to the calendar and participate more actively in online groups, but those hopes haven't yet materialized. Focused outreach may increase participation.
- **Developing a governance model**. Initially we operated without clearly defined roles, which allowed for flexibility but also meant that members didn't necessarily get recognized for their contributions and weren't tied to particular responsibilities. Moreover, we would like to develop a more coherent and transparent method for bringing new members onto the steering committee and rolling senior members off. In December of 2015, members of the steering committee came together for a day-long retreat at Rice University, where we discussed the need to spread out the work and ensure that all are recognized for their contributions and sketched out more defined roles. We are working on bylaws to formalize the organization's operations and are in the process of implementing plans made at the retreat. At the same time, many SC members are sympathetic to the lightweight organizational approach taken by NYCDH, which several Steering Committee members learned about during a recent presentation by Alex Gil and Kimon Keramidas at Digital Frontiers 2016. Focusing on partnerships and fostering communication across Texas DH organizations seems like a sound strategy for a small, all-volunteer organization like TxDHC.
- **Dealing with geographical distance:** Since Texas is the second largest state in the US, it can be difficult to connect people across such a significant distance. To cope with this distance, we have organized state-wide events and online gatherings, but we also hope to encourage members to use the website to promote events and activities within a particular city or region within Texas.

By working together, regional consortia can learn from each other, explore developing common infrastructure (as we have already benefited from CUNY's work on

Commons in a Box), and facilitate broader collaborations around research and teaching.

Florida Digital Humanities Consortium

Scot French

My presentation will explore the opportunities and challenges involved in creating a geographically extended, self-governing regional consortium (FLDH/Florida Digital Humanities Consortium) that depends, for its intellectual labor and technology support infrastructure, on academic institutional sponsorship while promoting free, open access to academic networks and DH resources.

Founded at THATCamp Florida 2014, FLDH's mission "is to provide a platform for studying and discussing digital tools, methods, and pedagogies as well as for educating teachers, faculty, and the public about the multiple, interdisciplinary ways humanities research and computing impact our world. It meets annually to identify issues of interest and to set goals for future collaboration and digital humanities research." At present the group has 12 institutional members, ranging from large research institutions to small liberal arts colleges: University of Florida (Gainesville), University of Central Florida (Orlando), Florida State University (Tallahassee), University of South Florida (Tampa), University of Miami, Florida International University (Miami), Rollins College (Winter Park), New College of Florida (Sarasota), Florida Southern College (Lakeland), Eckerd College (St. Petersburg), and the Florida Humanities Council. Inspired by NYCDH and the CUNY Academic Commons, group organizers adopted the Commons In a Box (CBOX) academic commons social networking platform as a virtual space in which to foster community and share resources.

An internal grant from the University of Central Florida funded an organizational meeting in Orlando at which representatives of participating schools and the Florida Humanities Council established an Executive Council with two representatives from each institution. A five-member elected Steering Committee drafted a mission statement and set of by-laws, established short- and long-term goals, and began planning for two major initiatives: Hosting HASTAC17 in Orlando (scheduled for November 2017) and submitting a proposal to host a southeast regional pre-conference workshop on visualization tools through the National Endowment for the Humanities' Institutes for Advanced Topics in the Humanities. Working on these initiatives has placed new demands on FLDH's leadership structure, and prompted a shift from volunteer initiative to a more active chair and working group/subcommittee organizational framework.

To date, all of our discussions about how best to build and organize our statewide/regional consortium have been internal to the group. We seek to expand the conversation to representatives of similar groups, particularly those using the CBOX platform developed by Matthew K. Gold and his project team at CUNY. Out of this exchange we hope to build an informal network of regional DH consortia for purposes of information sharing that can be expanded and formalized as needed.

This paper/presentation will raise issues of general interest to the DH community and of particular concern to members of the FLDH Executive Council (which I chair) and Steering Committee (on which I serve as a founding member).

- **Leadership/Self-Governance Structures**. What sorts of leadership/self-governance structures do regional DH consortia employ to ensure broad-based institutional representation, inclusion of diverse views, and transparency in setting group agendas and policies?
- **Academic Commons Activity**. What strategies or best practices might regional consortia adopt to build community and generate sustained activity on CBOX or other academic commons platforms? How can participating institutions most effectively contribute to the development of CBOX or similar platforms, with an eye toward enhanced functionality and customization?
- **Staffing/Program Support.** Is it feasible for participating institutions to "share" staff in support of mutually beneficial program and projects? What roles might digital humanities center staff, graduate research assistants, or postdoctoral research/ teaching fellows play in fostering a regionwide DH community and providing training for interested faculty and students? Is state, national, international, or foundation funding available in support of such regional collaborations?
- **Service/"Invisible Work."** To what extent, if at all, is the volunteer work of consortia officers (such as FLDH Executive Council/Steering Committee members) recognized as valued service within their respective institutions? Could regional consortia have a role in documenting and validating the contributions of active members seeking promotion and tenure or other paths to career advancement?
- **Geographic Distance/Virtual vs. Face-to-Face Meetings.** Are webinars and Google Hangouts an adequate substitute for face-to-face meetings and workshops? How might regional consortia spanning large geographic areas (such as Florida and Texas) create more funded opportunities for travel to consortia-sponsored conferences/meetings? What options are available both within and across member institutions for fundraising in support of travel?

As chair of the FLDH Executive Council and a member of the Steering Committee I will conduct an informal member survey to solicit other potential topics for discussion during this session. Ideas generated by this panel will be shared with the Executive Council and serve as the basis for discussion and action at an FLDH organizational meeting during the Orlando HASTAC conference in November 2017.

SD|DH: Building and Strengthening DH Teaching and Learning through a Regional Network

Erin Glass and Jessica Pressman

How can DH regional networks work to spread resources to underserved student populations, foster digital literacy and confidence in educators without DH institutional support, and strengthen local, humanistic forms of social advocacy? This presentation will focus on the development and future plans of the San Diego Digital Humanities (SD|DH) regional network, a collective comprised of faculty (and some staff and graduate students) from seven different local higher education institutions ranging from a R1 university to community colleges. While San Diego is not prominently known for humanities research—let alone DH—we recognize that its location on the border of Mexico, and the diverse student body of its institutions, make it an exceptionally rich site for socially-engaged, participatory forms of DH activity. By pooling resources, expertise, experience, perspectives, and moral support, SD|DH embraces the power of diversity as our chief value.

SD|DH has been working together for the last three years to share experiences and resources, plan events, apply for funding, and seek collegial support for our endeavors. We work from different institutional situations and relationships to digital humanities: a campus pursuing a grass-roots and ground-up approach to digital humanities research and teaching but lacking funding for institutional infrastructure, a campus administration wanting to implement digital humanities from a top-down structure, and a campus successfully implementing digital humanities projects within a specific department but lacking the leadership to build out from there. We have been successful in different ways at different campuses, but we now have a full-blown DH initiative at SDSU and an emerging program at USD, and we continue to use our individual campus efforts to bolster the region

Our first goal in forming SD|DH was to assess the barriers to implementing DH across a wide spectrum of institutions and diverse student populations, and develop work-around strategies for implementing DH by drawing upon the resources of multiple institutions within a single region. From the start, these efforts have been voluntary, but we were able to expand our efforts and visibility with the generous support of a National Endowment for the Humanities Level I Digital Start Up Grant for our project “Building and Strengthening Digital Humanities through a Regional Network.” This grant supported a series of workshops that focused on distributing DH to institutions and student populations usually left out of DH. It also helped encourage educators to explore new DH techniques for teaching within the safety net of a supportive community. Educators engaged in a wide range of DH methods such as text analysis and game design as well as DH-inspired techniques to creative re-imagine uses of everyday software and tools for pedagogical purposes. We met several times throughout the year to discuss challenges and successes, share expertise, and brainstorm new possible projects.

In this presentation, we will discuss our methods for developing and facilitating these workshops and pedagogical engagements as well as some of the relevant issues concerning coordination, labor, administrative support, technical resources, and transportation. We will also share lessons learned throughout our experience thus far and suggest protocols that could potentially be re-used and re-purposed by other institutions. Finally, we will argue for the need to collectively and creatively communicate the value of multi-institutional collaboration to administrators in order to increase administrators’ receptivity to these efforts going forward.

In addition to discussing projects already carried out, we would like to share our plans for moving forward. As part of deepening and expanding our community’s ties, SD|DH is working towards launching a multi-institutional digital commons powered by Commons in a Box (CBOX) hosted at UCSD. While some SD|DH campuses could ostensibly run and host their own CBOX we have decided to experiment with creating a single commons for our community so that all participating SD|DH institutions can take advantage of the many affordances of CBOX. We are planning to use this Commons in multiple ways. First, we will use it to facilitate communications for the SD|DH group as a list serv, events calendar, member directory, and showcase of SD|DH projects. However, we will also invite all students, faculty, and staff of SD|DH institutions to use the Commons for networking across institutions, building websites related to research and teaching (for any discipline), and creating interest groups that cut across disciplines and institutions. Finally, we also hope to implement Social Paper, a collaborative writing tool developed at The CUNY Graduate Center for CBOX through the generous support of a NEH Digital Start Up Grant. We are currently discussing the possibility of designing multi-institutional synchronous courses that will engage students in thinking through the role of collaboration in their respective disciplines by collaborating with students outside of their institutions through Social Paper.

Bibliography

Alexander, B. and Frost Davis, R. (2012) "Should Liberal Arts Campuses Do Digital Humanities? Process and Products in the Small College World." Debates in the Digital Humanities, Ed. Matthew K. Gold. Minneapolis: University of Minnesota Press

Spiro, L. (2014) “Creating the Texas Digital Humanities Consortium.” Digital Scholarship in the Humanities. April 23, 2014.

https://digitalscholarship.wordpress.com/2014/04/23/creating-the-texas-digital-humanities-conso rtium/

Theibault, J. (2016) "Regional Digital Humanities Consortia: An Emerging Formalization of Informal Network Ties? [Poster]." In Digital Humanities 2016: Conference Abstracts, 902–3. Kraków: Jagiellonian University & Pedagogical University. http://dh2016.adho.org/abstracts/176.

Beyond Access: Critical Catalog Constructions

Molly Hardy
mhardy@mwa.org
American Antiquarian Society, United States of America

Dawn Childress
dchildress@library.ucla.edu
UC Los Angeles, United States of America

Paige Morgan
paige.c.morgan@gmail.com
University of Miami, United States of America

This panel explores the use of digital, rare book catalogs as platforms for collaboration and as sources of data to uncover patterns of book production and to offer new insights into the sociology of texts. Extractions from and additions to bibliographic data extend the catalog beyond its original use as a point of access and discovery. The use of catalogs as sources of bibliographic big data as well as the use of platforms that enable bibliographic record annotation allow us to reimagine the catalog's trajectories of creation and utility. Considering the genealogy of the catalog, the panel will examine how research practices have both informed and been informed by catalogs since their inception and how the work of bibliography is reimagined in the catalog. Focusing on how limitations and aporia in the catalog can lead to critical making in the digital age, the panel will show how digital uses of the catalog currently enable mindful interrogation of catalog data and catalog making as well as consider possibilities for expanded use of rare book catalog data in the future. This consideration of the rare book catalog as a digital humanities project invites reassessment of legacy information architecture as well as the many hands that built the bibliographic structures on which so much of the work of the digital humanities rests.

In "Towards Speculative Catalogs," Dawn Childress will open with a discussion of the transformative promise of the digital as we reconstruct catalogs in new forms and formats. To provide context for critical catalog constructions more broadly, Childress will highlight how bibliographies and catalogs have served as source material for research beyond that of points of access in both pre- and post-DH contexts, as well as consider how digital humanities use cases might differ from more analog approaches, whether qualitative or quantitative. In addressing these questions, Childress will explore the promise of applying current and emerging tools and standards (such as linked data, IIIF, PCDM, etc.) to the practice of interrogating bibliographic and catalog data. Childress will suggest how we might leverage these systems to record and analyze lacunae, erasures, and bias in capturing the bibliographic record and, drawing on Bethany Nowviskie's notion of speculative collections, how these systems might support active reframing and interrogation by users.

In "'The Technology of Shared Cataloging': A Retrospective," Molly O'Hagan Hardy will build on Childress's remarks through a close look at the creation and re-creation of two rare book union catalogs: the English Short Title Catalog (ESTC) and the North American Imprints Program (NAIP). In 1981, in a Bibliographic Society of America Symposium from which the title of Hardy's paper takes its name, William Todd wrote, "Perhaps we do not yet fully appreciate the situation, now rapidly materializing, whereby computers converse with each other in any mode, while the rest of us, mere mortals, stand mute before them." Remarks like this, which abound in the excitement and trepidation expressed during the emergence of these rare book union catalogs, echo a similar exuberance and hesitancy around the transformation from MARC to linked data models. Examining what "machine readable" meant then and means now, Hardy will draw parallels between the current conversation around BIBFRAME and other such initiatives and those early efforts led by Robin Alston and Marcus McCorison to amass large amounts of special collections' catalog data. She will then examine what it was these catalogs set out to capture and in what ways this work is being reimagined in the linked open data environment. She will do this through a close look at the American Antiquarian Society's Printers' File linked data project and its reliance on LCNAF and VIAF to merge MARC data with BIBFRAME. Ultimately, she will consider how such initiatives necessitate the reimagining of library and scholarly work, so those working on both sides of the reference desk are not left to "stand mute" before their creations. Hardy will point to examples of innovative uses of rare book catalog data for digital humanities projects. Such uses, Hardy will show, not only prove efficient and effective means of generating data, but they also productively unearth biases inherent in any information system.

Paige Morgan will conclude our presentations with "Searching For Common Ground: Modeling Bibliographic Data in Library and DH Contexts," in which she will present the results of a survey of data use in DH projects examining the projects' use and presentation of library data (including bibliographic data and data presented in digital collections platforms). This survey will focus on

- whether each project includes bibliographic references;
- how much granular detail any bibliographic references include,
- whether or not such data is presented in the format of a specific model, and
- the presence or absence of links to specific copies, whether in library or digital archive catalogs (such as HathiTrust, the ESTC, paywalled collections (ECCO, EEBO, etc.), or websites like Google Books and Project Gutenberg).

Gathering this data will allow Morgan to look for common priorities in the project creators' use of library data; and to identify some of the assumptions that digital humanities has made about libraries and the bibliographic information that they produce. She will use the DH project survey data as the basis for a comparison with bibliographic ontologies (such as FRBRoo, BIBFRAME, Schema.org, and BIBO), and literature on challenges and best practices in bibliographic data modeling. Morgan will look for intersections, missed connections, and opportunities between digital humanities and cataloging work; and will consider how assumptions made in digital humanities elide the complexity and ongoing negotiations in the production of bibliographic data. In these projects, what drives the decisions to model (or not model) bibliographic data? How do the priorities of DH practitioners differ from those of library-based data creators? The increasing development of off-the-shelf tools, and the gradual growth of infrastructure for learning digital humanities skills and accessing data means that in many ways, it is possible for DHers, both new and experienced, to do more than before. However, increased access to materials and resources does not mean that the efforts of DH and library data communities will automatically complement each other. In FRBR, Before & After, Karen Coyle observes that library personnel working with data "have made little change in our approach to subject analysis in the last half-century, possibly because there isn't a clear direction for improving this aspect of our work." This paper will argue that similar ambiguities exist around the use of bibliographic data in DH projects, and that the apparent common ground of bibliographic data use is more complex than it appears.

Decolonizing Methodologies: Recovery and Access Amidst the Ruins

Christy Hyman
christy@huskers.unl.edu
University of Nebraska-Lincoln
United States of America

Anelise Shrout
ashrout@fullerton.edu
California State University-Fullerton, United States of America

Kathryn Kaczmarek-Frew
kkaczmar@umd.edu
University of Maryland, United States of America

Hilary Green
hngreen1@ua.edu
University of Alabama, United States of America

Nishani Frazier
frazien@miamioh.edu
Miami University, United States of America

Brian Rosenblum
brianlee@ku.edu
University of Kansas, United States of America

"It was as if these matters of objective and hard science provided an oasis for folks who did not want to clutter sharp, disciplined, methodical philosophy with considerations of gender, race and class-determined facts of life" (Earhart 2012, pg. 5).

"There is something truly soul-destroying in the repeated discursive erasure of vulnerable identities whether in the media, on the street, in the classroom, or in the legislative chamber" (Wright 2015, 264).

Recent instances involving the desecration of hallowed structures erected in remembrance of turbulent periods in the U.S. solidify the need for a call to action to enact historical recovery amidst the "ruins." The ruins include the marginalization of subaltern agents within grand historical narratives, the propensity of the academy to view historical purveyors of intellectual thought as overwhelmingly European, and the invisibility of subaltern agents in traditional and digital archives. Sites of subaltern experience drift in and out of the historical record often leaving no tangible signifiers to how these events inform the racialized, gendered, and class based systems of power which influenced many of the negotiations that subaltern agents contended with in initiating action and decision-making in an oppressive world. Historical recovery in the digital humanities play a vital role in enabling users with access to new knowledge as well as helping to shape meaning and interpretation of the cataclysmic phenomena embedded within its mediums. Anelise Shrout's call for **history from below** is essential to creating digital content which enables accessibility and users potentiality of increased awareness of the historical underpinnings related to structural inequality- Shrout rightly declares that digital methodologies are ideally suited for these interventions (Shrout 2016).

This panel offers several approaches to historical recovery through showcasing digital projects which utilize technologies designed to foreground subaltern histories while preserving contextual integrity. We bring together three approaches- geospatial, alternate reality gaming, and online digital collections for a conversation that will highlight the methods and values involved in the digitization, visualization, user interaction and theoretical analysis of content. Our panel consists of scholars who've developed interdisciplinary approaches in their area of inquiry emphasizing historical recovery as a focal point of their research: Anelise Shrout's presentation explores the Digital Almshouse Project and argues for the importance of digital humanists bringing together new archival methods with the old while critically reading archival silences as opportunities for further recovery and analysis. Christy Hyman's presentation highlights the ways that GIS recovers subaltern experience through the struggles of enslaved runaways in antebellum eastern North Carolina using the digital narrative The Oak of Jerusalem: Flight, Refuge, and Reconnaissance in the Great Dismal Swamp Region. Hilary Green's paper examines digital collections alongside pedagogical approaches that introduce students to centering historically subaltern actors using intersectional strategies and inclusive pedagogies. Kathryn Kaczmarek-Frew illuminates opportunities available with alternate reality games in building adolescent's historical knowledge by showcasing The Tessera, a game inspiring teens to take on the roles of real life scientists, programmers, and artists as they solve conceptual challenges while immersing themselves in a storyworld. Finally, Nishani Frazier discusses the conflictual relationship between the academy and the digital, the changing role of the scholar/expert, the importance of community in historical production, and the balance needed between recovery and shared authority.

In doing so, this panel promises to stimulate a lively and engaging discussion that will be informative to digital humanists of many fields. By presenting content and methods that connect subaltern humanity to recovery methods within the digital this panel helps forge the way for future work in gaining access to a more public facing scholarship that foregrounds the marginalized stories, agents, and actors into the digital humanities.

All panelists have agreed to participate.

Rehumanizing Data: Digital Methods Against Archival Violence

Anelise Shrout

Some of the most interesting historical subjects left only the faintest traces in the archive (Shrout 2016). When people of color, enslaved people, immigrants or women appear at all, it is often as commodified figures, ancillaries to those in power, or as undifferentiated masses of undesirable bodies. Scholars have recently begun to note that the archival silencing of these marginalized voices is not merely oversight, but rather, as Marisa Fuentes argues, a kind of epistemic violence. (Fuentes 2016). The absence of these subjects, these scholars note, is a feature, rather than a bug.

This presentation proposes a set of digital humanities practices that explicitly work against the structures that kept those voices silent in the first place. It uses Digital Almshouse Project, a dataset of Irish immigrant who emigrated to North America to the port of New York in the 1840s. These men and women subsequently fell ill, became destitute, or were simply seen to be incompatible with New York's public spaces. They were identified by public officials and sent (often forcibly) to New York's Bellevue Almshouse – the only public health site that offered social services to people fleeing famine. In the process they were transformed; their diseases (a category which included the social maladies "recent emigrant" and "destitute"), their children (who they would often be separated from inside the almshouse) and the spaces of confinement to which they were sent transmuted from individual experience into bureaucratic data.

This presentation uses the Digital Almshouse Project to rehumanize nineteenth-century immigrants. In doing so, it illustrates the ways in which scholars can bring "big data" analysis to bear on silenced voices; can use data visualization to resurrect archival ghosts; and can deploy quantitative analysis to give weight to historical actors long since written out of history. It asks that data-driven digital humanists pay particular attention to historical actors whose bodies were quantified and commodified, and who were in the process stripped of markers of humanity. Finally, it highlights historical bureaucratic spaces in which immigrants were able to enact agency, while also laying bare the ways in which the carceral state has used data to dehumanize marginalized subjects.

The methods discussed in this presentation have roots in "history from below," the "new social history" and cliometrics of the 1970s and 1980 (Gallman 1977, Shammas 1977, Darnton, 1984). Together, these methods sought to resurrect the stories of "people with no history," by reconstructing their quotidian experiences. However, the subjects of these studies tended to only be relatively marginalized – they included (white, male) artisans, (white, male) voters, and (white) women property holders. There have been various resurgences and developments in these methods in the intervening four decades. These include practices of reading archives "against the grain" to get at the unstated assumptions that historical actors made about those they held power over. They also include theoretical approaches that advocate the reading of silences to understand those whose voices were intentionally obscured by official recorders and gatekeepers (Trouillot 1995, Bastian 2003, Drake 2016). This presentation argues that we must bring together new archival methods and the old cliometrics; that we must critically read archival silences while simultaneously using regression models to coerce meaning from our data; that we should think about the motives of historical data creators, while we also use dataviz to trace historical actors' pathways through hostile spaces. This presentation aims to not only theorize digital approaches to marginalized people, but to present a concrete toolkit for beginning to center the voices and experiences of those marginalized historical subjects.

GIS as a Phenomenological Bridge to Experience: Deep Mapping the Enslaved Runaways of Eastern North Carolina

Christy Hyman

My paper examines African American efforts toward cultural and political assertion in geographies of domination, specifically the Great Dismal Swamp area and adjacent communities in North Carolina during the long nineteenth century. Building on the work of archaeologists, historians, and novelists, my presentation reimagines space as it relates to the experiences of enslaved runaways and laborers who were exposed to the Great Dismal Swamp. I argue that the proximity of the Great Dismal Swamp increased the incidence of flight despite the danger and overall inhospitable nature of that landscape. In examining African Americans who navigated the landscapes of trauma inherent in the institution of slavery within the eastern North Carolina region, this study sheds light on the common values, aspirations, culture, and economic systems of a people relegated to the margins of society. This investigation into the spatial dimensions of flight within the Great Dismal Swamp, with the use of historical primary sources, generates new knowledge about the enslaved experience. This approach highlights and analyzes space itself with a view to uncovering the social relationships embedded in it.

Focusing on enslaved runaways' proximity to waterways and the natural environment within the Great Dismal Swamp, this project uses Esri Story Maps to tell the digital narrative of The Oak of Jerusalem: Flight Refuge and Reconnaissance in the Great Dismal Swamp Region, a project containing maps with narrative text, images, and multimedia content. The data contained within this digital narrative was gathered from nineteenth-century North Carolina newspapers such as the Cape Fear Mercury, Carolina Observer, Edenton Gazette, and many others which provided information on the destinations and distributions of enslaved runaways, including frequency distribution of enslaved runaways by month of disappearance and gender specificity. In terms of the Great Dismal Swamp landscape, content will include remote sensing data, forest dynamics of the swamp, soil characteristics, and animal species such as amphibians, reptiles and mammals. Taking these physical properties of space together with the historical descriptions of enslaved runaway ads and cross-referencing this information with textual descriptions in the enslaved narratives written by Moses Grandy and Harriet Jacobs, a complete reality of what was at stake for an enslaved person to run away emerges. Because these seemingly disparate elements of space (and the social relationships embedded within) are scattered across the archival landscape, my project brings all these things together in one setting.

Geospatial technologies are central to this study because of the ways that GIS can be a phenomenological bridge between subaltern experience and measurable qualities of the Great Dismal Swamp landscape. Using geoprocessing methods employed with ArcGis tools a cost surface was developed to investigate the ways enslaved people appropriated the landscape into areas of refuge and reconnaissance and eventual escape. A cost surface is a geospatial technique based on measuring the effort it takes to cross a landscape area. Data contained within a DEM (Digital Elevation Model) representing the Great Dismal Swamp landscape allows for this form of spatial analysis. The cost surface can thus be used to find the path between identified points along the surface with the least amount of effort. In measuring "cost" for enslaved people during the antebellum era four sets of variables were attained:

1. Cumulative distance- total distance to travel from one place to another
2. Duration or travel time- total time to travel from one place to another
3. Energetic expenditures or calories- total calories expended to travel from one place to another
4. Experiential/Cultural- abstract measures such as visibility between sites, social distance, and spheres of influence.

Through Moses Grandy and Harriet Jacob's dynamic reminiscences, the extant data on enslaved runaways near the Great Dismal Swamp, and the physical properties of space within the Great Dismal Swamp landscape the shadowy foundations of how the Great Dismal Swamp functioned as a quasi-dystopian landscape for enslaved people is recovered. Ultimately the swamp embodied the promise of freedom as a site of possible refuge, but also the horrors of the slavery regime itself.

Disrupting the Archive: Intersectionality and the Integration of the Digital Humanities in the Classroom

Hilary Green

In today's classroom, gender and race scholars strive to provide students with a breadth of scholarship that represent the social, cultural, and political experiences that define both the American and global experiences. For scholar-activists working at large southern PWIs, the politicized tensions present in society today stemming from difficult racial pasts, and potential efforts at truth and reconciliation, all compel us to reconsider what sources will most impact students' understanding of these difficult subjects. To effectively educate students, instructors do well to expose them to diverse historical documents, assignments, and approaches. Digital humanities offer new opportunities and challenges as eloquently expressed by Tara McPhersons' "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation." This conference presentation will explore how to use engaging assignments (spatial analysis projects using Carto and documentary history projects) and digital archive collections to introduce students to a variety of sources using intersectional approaches and inclusive pedagogies that are used to critically engage students from diverse backgrounds.

Digital archive collections offer scholars a valuable service by providing an online platform to publish digital archives. Managing content is no easy task and scholars as well as institutions have to find resources and time to develop their sites in a way that encourages curiosity, usability, and scholarly utility.

Colored Conventions is an excellent archival addition to the digital humanities. Colored Convention's purpose deserves quoting in full:

> "This project seeks to not only learn about the lives of these mostly male delegates, the places where they met and the social networks that they created but to also account for the crucial work done by Black women in the broader social networks that made these conventions possible."

ColoredConventions.org endeavors to transform teaching and learning about this historic collective organizing effort—and about the many leaders and places involved in it—bringing them to digital life for a new generation of undergraduate and graduate students and researchers across disciplines, for high school teachers, and for community members interested in the history of church, educational and entrepreneurial engagement" (Colored Convention, 2015).

As evidenced by Colored Convention's purpose their aims are to reach a broad audience which refers to the public humanities focus of the project. The site has teaching aids, curriculum guides, as well as the primary source documents that have been recovered from relative obscurity. What Colored Conventions does is recover an entire archive of action performed by antebellum era African Americans taking initiative to move closer to freedom.

In the context of pedagogy, intersectionality asks students to think beyond narrow binaries that overdetermine contemporary discourse on identity and power (Risam, 2016). For particularly vulnerable populations who have historically experienced economic disadvantages the disciplines under the humanities does not at first glance offer a high upward mobility. How then can the humanities compete and attract a diverse array of students as well as engage scholars toward its humanistic offerings? In engaging digital archive collections centering marginalized people's histories it becomes necessary to interrogate the archive, recover subaltern experience, and reimagine new approaches to historical interpretation. In doing these things fresh perspectives unfold as the humanities discipline as a whole is invigorated with scholarly interest. Students and scholars alike find new purpose in the humanities. These new questions coupled with the innovation in toolmaking within the digital invite a wealth of promise through intersectional approaches and inclusive pedagogies.

Casting our Lights Before: Immersion as Access in The Tessera ARG

Kathryn Kaczmarek-Frew

"They say that coming events cast their shadows before. May they not sometimes cast their lights before?" -- Ada Lovelace

Computer science and the technology industry are often regarded as the bastion of white men. However, there is a movement in popular culture to recover and make visible the contributions of women and minorities, from the major motion picture Hidden Figures to the AMC series Halt and Catch Fire. Still more active methods can be used to connect today's youth to the realities of the past and the promise of the future. Alternate reality games provide a unique opportunity to combine historical knowledge and skill-building in a way that pervades into daily life. In my presentation, I will show how our alternate reality game The Tessera invites teen players to develop a more nuanced understanding of computer history while demonstrating that they already possess the computational thinking skills to make a future in the industry accessible to them.

Developed as a collaboration between the University of Maryland, Brigham Young University, and the Computer History Museum under a grant from the National Science Foundation, The Tessera ARG uses the genre of Victorian gothic realism and the trope of haunting to think about the influence of the past on the present. The narrative focuses on the ghost of Ada Lovelace recruiting teenage players, particularly those from demographics underrepresented in STEM disciplines, to join her secret society known as The Tessera in their fight against an even more mysterious destructive force known as S. Racing against time, players complete a series of collaborative puzzles emphasizing computational thinking skills such as pattern recognition, decomposition (breaking down a problem into smaller steps), algorithmic design, and data representation. By emphasizing the STEAM values of innovation and creativity as the key to progress, The Tessera helps players recognize that they already do computational thinking and that they can have a future in the technology industry.

Like many ARGs, The Tessera distributes its narrative across several media, presenting different opportunities for immersion. The Tessera: Ghostly Tracks is a real-life experience in the Computer History Museum, where classes will divide into groups to use digital clues and physical artifacts in the Revolution exhibit to solve computational thinking puzzles and discover which famous figures from computer history are haunting the museum. Players will experience a highly embodied sense of immersion as they move around the exhibit and manipulate props to solve the puzzles, such as decoding the message etched on the top of a chessboard or using a code to open the six padlocks on an "enchanted" crate, releasing the ghosts from the museum. In contrast, The Tessera: Light in the Shadows is an online game in which Ada guides players through the ruins of a decaying Victorian pub into the Tessera stronghold of Horsley Towers to create new innovations inspired by key inventions from history (such as the telegraph, the metric system, and Wi-Fi) to thwart S's destruction. The online game requires imaginative recentering to immerse the player within the storyworld, aided by the use of first person perspective, tension-inducing time limits, and the invitation to write themselves into the narrative through the public-facing wiki (Ryan 1992). While the goal is for players' personal investment in the Tessera's struggle to translate into learning, I'm interested in studying how the different avenues of immersion in the two experiences influence players, not only in their interest in the game content but also their self-perception of their computational thinking skills. This research will allow future game designers to think about how different kinds of simulations (physical, digital, augmented, and virtual) can be used for historical reconstruction to help players best access the realities of buried pasts.

Power to the People: Recovery and Shared Authority in Civil Right Movement Digital Work

Nishani Frazier

African American Studies, public history, and oral history methodologies play central roles in the recovery of fairly unknown aspects of civil rights movement history. However, this merger also required a philosophical embrace of community engagement, power sharing, and open access during historical production. Though these ideals may fit within digital humanities, historical practice among academics is at odds with this approach. This talk discusses the conflictual relationship between the academic and digital, the changing role of the scholar/expert, the importance of community in historical production, and the balance needed between recovery and shared authority.

Theorizing African American Public History Through the Digital Humanities

Expanding public understanding about the history of the Congress on Racial Equality community organization and economic development through access to primary sources, teacher resources, and mapping is the resource cornerstone within this project. It acts to recover a little known aspect about CORE and its relationship to racial uplift and economic development while providing access for future scholars and other interested persons. This process is a model for intersecting African American scholarship and digital humanities.

The digital representation Harambee City illustrates this philosophy through:

1. Broad accessibility
2. Providing a second layer of learning through open access to primary sources and teacher lesson plans
3. Knowledge sharing - commentary/exchange on the website
4. Community organization training

The aim of this project is to promulgate further inquiry within historical study of the Civil Rights Movement as well as contemporary interventions concerning social justice issues (Frazier 2016). Following on Jessica Marie Johnson's call to action regarding the urgent need for the digital humanities to engage with society's "marginalized or discriminated against" this project endeavors to fulfill historical as well as justice imperatives needed within the digital humanities by examining civil rights movement history not commonly centered (Dinsman 2016).

Notes

1. The Old Ashburn School, an historic African American schoolhouse in Loudoun County, Virginia was vandalized with swastikas on October 1, 2016, another incident occurred in Tallahatchie County, Mississippi where the memorial sign of murdered African American teenager Emmett Till was found riddled with bullets on October 15, 2016. As far as the Pacific Northwest in Portland, Oregon in September of 2016 the statue of Vera Katz, a progressive politician and former mayor of the city was sprayed hateful epithets. Katz is of Jewish Menshevik origin and her family fled Nazi controlled Germany in 1933.

Bibliography

Bastian, J.A. (2003). Owning Memory: How a Caribbean Community Lost its Archives and Found its History. Westport: Libraries Unlimited.

Colored Conventions.org. (2016). Omeka RSS. http://coloredconventions.org/ (accessed August 10, 2016).
Darnton, R. (2009). The Great Cat massacre: and Other Episodes in French Cultural History, New York: Basic Books.
Dinsman, M. (2016) The Digital in the Humanities: An Interview with Jessica Marie Johnson. LA Review of Books. https://lareviewofbooks.org/article/digital-humanities-interview-jessica-marie-johnson/ (accessed July 24, 2016).
Drake, J. (2016) Archives: Towards Belonging and Believing. https://medium.com/on-archivy/liberatory-archives-towards-belonging-and-believing-part-1-d26aaeb0edd1#.fht1cp2mj (accessed March 26, 2017
Earhart, A. (2012). Can Information be Unfettered? Race and the New Digital Humanities Canon. Debates in the Digital Humanities. U. of Minnesota P. Matthew K. Gold, (Ed) http://dhdebates.gc.cuny.edu/debates/text/16 (accessed October 20, 2016).
Frazier, N. (2016) Harambee City http://harambeecity.lib.miamioh.edu/use-and-restrictions (accessed 10/6/2016).
Fuentes, M. (2016) Dispossessed Lives: Enslaved Women, Violence, and the Archive. Philadelphia: University of Pennsylvania Press
Gallman, R.E. (1977). Some Notes on the New Social History. The Journal of Economic History, 37(01): 3–12.
Risam, R. (2016) Intersectionality. Digital Pedagogy in the Humanities: Concepts, Models, and Experiments: Modern Language Association https://digitalpedagogy.commons.mla.org/keywords/intersectionality/ (accessed October 30, 2016).
Ryan, M.-L., (1992). Possible worlds, Artificial Intelligence, and Narrative Theory, Bloomington: Indiana Univ. Press.
Shammas, C., (1977). The Determinants of Personal Wealth in Seventeenth-Century England and America. The Journal of Economic History. 37(03): 675–689.
Shrout, A. (2016) "Digital History from Below: A Call to Action." Digital Humanities 2016 Conference Abstracts. dh2016.adho.org/abstracts/347 (accessed October 12, 2016).
Trouillot, M.R. (1995). Silencing the Past: Power and the Production of History. Boston, MA: Beacon Press.
Wright, M. (2015). Physics of Blackness: Beyond the Middle Passage Epistemology. Minneapolis: University of Minnesota Press.

Visualizing Futures of Networks in Digital Humanities Research

Micki Kaufman
mickikaufman@gmail.com
The Graduate Center, CUNY, United States of America

Zoe LeBlanc
zoe.leblanc@vanderbilt.edu
Vanderbilt University, United States of America

Matthew Lincoln
mlincoln@getty.edu
Getty Research Institute, United States of America

Yannick Rochat
yannick.rochat@unil.ch
University of Lausanne, Switzerland

Scott B. Weingart
scottbot@cmu.edu
Carnegie Mellon University, United States of America

Introduction

Papers including the topic "Networks, Relationships, Graphs" have comprised roughly 10% of submissions to ADHO's annual conference for the past 4 years - a sizable portion, to be sure, but one that has remained roughly consistent in that time (Weingart, 2015). "Networks" are, in the abstract, familiar to humanities scholars devoted to studying complex relationships. This potential is alluring, but advanced network analytical techniques are challenging to implement and interpret. And overly complex visualizations have attracted derogation from some scholars, deriding visually-impressive but uninterpretable graphs as "hairballs."

This roundtable will take up crucial questions: What kinds of data, questions and interpretive techniques are appropriate for network analysis? How does the disciplinary skillset of the humanist researcher determine, enable or limit effective network analysis? To what extent does the use of data visualization serve to surface, or submerge, essential knowledge about the data? How should scholars in the digital humanities navigate the intense methodological demands of network science? How should such scholarship be evaluated, peer-reviewed, taught, and studied? In the face of these many challenges, what are the futures of networks in DH?

Network Sources / Network Evidence

Why transform our research sources into networks? For some projects, the simple reframing of evidence as a network visualization provides a sufficiently novel perspective to pose more precise research questions and to isolate specific avenues for more research. For research fundamentally about network structures and dynamics, more advanced techniques, including simulation and quantitative hypothesis testing, are required to produce valuable results.

Which path to take may depend on one's sources. Some sources are naturally transformed to networks: correspondence from one individual to another (Winterer, 2012; Ahnert and Ahnert, 2015), for example, or kinship relations. (Jenkins et al., 2013) But less obvious sources may also be seen as networks, such as characters co-occurring in a plot, or documents connected by shared topics. The abstracting and filtering effect of network analysis can also be powerfully applied to illuminate how sources themselves interact to construct knowledge of subjects (Kim, 2013).

How can we encourage more creative thinking about transforming sources (from collections, archives, texts, objects, and more) into networks? When is “basic” visualization productive by itself? Where are complex methods like agent-based simulation or predictive modeling best used? How can network analysis be used to illuminate power imbalances within the scholarly infrastructure? What are strategies for dealing with known unknowns (and unknown unknowns!) in network research, and how can we visualize these missing data?

Disciplinary relationships: Complexity science, humanities, and DH

Examples of the “network” or “graph” idiom, whether actually visualized or merely referenced within a text, can be found in citations well predating modern-day tools for network analysis. They are numerous in sociology (Freeman, 2004), but also in the history of art (Barr Jr, 1936), anthropology (Gell, 1998; Hage & Harary, 1983; Foster, 1969), geography (Bertin, 1967), and economics (Koenig et al., 1979), among others. The idea of the network is a seductive one for humanists who wish to study the multilayered web of interactions between any number of agents (authors, texts, readers, artists, artworks, viewers, patrons), in order to discern how those interactions produce structure and meaning all their own. To do so, however, scholars must grapple with guidelines for expressing assumptions, formulating hypotheses, and gathering and testing evidence using a language of network theory and sociology that can seem alien, if not inimical (Galloway and Thacker, 2007). How have humanities scholars navigated this challenge when using network analysis?

Compounding this effort is the rapid expansion of network and complexity science in its own right. This rapid evolution challenges humanists who would adopt some of these methods for their research. Can a single scholar can find their way without formal partnership with a collaborating network scientist? This raises issues particularly for peer review: How are these papers evaluated between their methodological and their content disciplines?

Network visualization

As with its determination and preparation, visualizing humanities network data in a comprehensible manner is an inherently interdisciplinary task that requires a knowledge of the academic domain, rigorous archival and data management work, and an effective engagement with visual design practices. The proliferating use of visualization tools to represent network data in the digital humanities demonstrate both the potential and the difficulty of this undertaking. The immense complexities of the human connections that network visualizations represent and the probabilistic mathematics that distribute its nodes combine to confound and defy consistent interpretation. Basic technical constraints of dimension, visual design traditions, and a relentless drive for legibility all further reduce, constrain, or even determine the possible interpretations of a dataset from a diagram.

What can humanities researchers engaged in the active process of network visualization do to make informed and effective computational, interpretive, aesthetic and practical decisions? In what cases is the beleaguered “hairball” still a productive or generative approach, in spite of the difficulty it can pose to interpretation? What other alternatives exist? How can the tools, design traditions and/or algorithms currently in use, as well as the introduction of new approaches, dimensions and technologies enhance the power of a network visualization to express and communicate essential understandings about humanities datasets?

Networks and Interactivity

How could new dynamic interactions with network visualization help us better understand and explore our data? With the rise of data journalism and in-browser apps, network visualizations are increasingly interactive, using animations and dynamic features to visualize additional dimensions. Such interactivity can help further an argument, and encourage the user to engage with the data. But, how sustainable and accessible are these visualizations? The long-term viability of these network visualizations depends on continued support, from updating code libraries to adapting to new browser requirements. Moreover, interactivity can be too demanding for slow internet connections, while also complicating workflows for both print and online publication. Added interactivity may also foreground style over substantive engagement with research questions.

What is the relationship of these interactive graphs to their textual explications? How can we design interactive visualizations for multiple modalities and bandwidths? How can digital humanists determine when interactivity is furthering their network analysis? How might interactive network analysis leverage the insights of social annotation tools to analyze metadata on users’ interactions with network visualizations, or utilize more immersive digital experiences, such as virtual or augmented reality?

Access to network methods and tools

All of these challenges intersect with how we teach network analysis and how the scientists teach themselves. The algorithmic transformations of network analysis are not easily accessible, and present a major barrier, particularly to those without any background in data analysis or programming. Those network analysis tools that are accessible to newcomers - and are thus frequently taught in short-term DH workshops - privilege the visualization of networks while largely concealing the behind-the-scenes work of network metrics calculation.

As with computational text analysis, it is simply beyond the scope of graduate programs in the humanities to take on complete responsibility for training its students in network analysis methods (Underwood, 2014). What strategies in mainstreaming computational textual analysis within DH (e.g. the emergence of dedicated "text labs" at several institutions) could be used to produce more substantive work in DH? What failures should be avoided? Where does network analysis in DH diverge so much from computational text analysis that entirely new strategies need to be considered? Moreover, how can practitioners of network analysis in DH make their research understandable and accessible to a larger audience?

Bibliography

Ahnert, R., & Ahnert, S. E. (2015). Protestant Letter Networks in the Reign of Mary I: A Quantitative Approach. ELH, 82(1), pp. 1–33.

Barr Jr., A. H. (1936). Cubism and Abstract Art. New York, Museum of Modern Art. Reprint, 1966.

Bertin, J. (1967). Semiology of Graphics: Diagrams, Networks, Maps. Esri Press. Reprint, 2010

Foster, G. M. (1969). Godparents and social networks in Tzintzuntzan. Southwestern Journal of Anthropology, pp. 261–278.

Freeman, L. C. (2004). The development of social network analysis. A Study in the Sociology of Science.

Galloway, A. R. and Thacker, E. (2007). The Exploit: A Theory of Networks. Electronic Mediations 21. Minneapolis, University of Minnesota Press.

Gell, A. (1998). Art and Agency: An Anthropological Theory. New York, Clarendon Press, p. 235.

Hage, P. and Harary, F. (1983). Structural Models in Anthropology. Cambridge University Press.

Jenkins, N., Andrews, A., Meeks, E., Grossner, K. and Murray, S. (2013) Kindred Britain. http://kindred.stanford.edu (accessed on Oct. 31, 2016).

Kim, D. J. (2013). "Data-Izing" the Images: Process and Prototype. in Performing Archive: Curtis + "the Vanishing Race," ed. Jacqueline Wernimont. http://scalar.usc.edu/works/performingarchive/data-izing-the-photos (accessed on Oct. 31, 2016).

Koenig, T., Gogel, R. and Sonquist, J. (1979). Models of the Significance of Interlocking Corporate Directorates. American Journal of Economics and Sociology, 38(2), 173-186.

Underwood, T. (2014). How Much DH Can We Fit in a Literature Department?. https://tedunderwood.com/2014/03/18/how-much-dh-can-we-fit-in-a-literature-department/ (accessed on Oct. 31, 2016).

Weingart, S. B. (2015). Submissions to DH2016 (Pt. 1). http://scottbot.net/submissions-to-dh2016-pt-1/ (accessed on Oct. 31, 2016).

Winterer, C. (2012). Where is America in the Republic of Letters?. Modern intellectual history, 9(03), pp. 597–623.

Open, Shareable, Reproducible Workflows for the Digital Humanities: The Case of the 4Humanities.org 'WhatEvery1Says' Project

Alan Liu
ayliu@english.ucsb.edu
UC Santa Barbara, United States of America

Scott Kleinman
scott.kleinman@csun.edu
CSU Northridge, United States of America

Jeremy Douglass
jeremydouglass@english.ucsb.edu
UC Santa Barbara, United States of America

Lindsay Thomas
lindsaythomas@miami.edu
University of Miami, United States of America

Ashley Champagne
ashleychampagne@umail.ucsb.edu
UC Santa Barbara, United States of America

Jamal Russell
jamalsrussell@umail.ucsb.edu
UC Santa Barbara, United States of America

Introduction

This panel reports on the open, shareable, and reproducible workflow methodology for digital humanities research developed by the 4Humanities.org "WhatEvery1Says" (WE1S) project. WE1S is topic modeling a large corpus of articles related to the humanities in newspapers, magazines, and other media sources in the U.S., U.K., and Canada from 1981 on. While the panel presents WE1S's conceptual goals and prototype experiments in using outcomes in humanities advocacy, its focus is on the technical and interpretive workflow developed by the project for humanities-oriented data work. WE1S's *manifest system* for data provenance and workflow management, its *virtual workspace manager* for integrated, containerized data manipulation and processing, and its *interpretation protocol* for how humans read topic models suggest a generalizable open approach based not on particular technologies and methods but on annotated methods. Moreover, there is a philosophical fit between such an approach and the public-facing goals of the WE1S project. WE1S is about opening public culture to view through analytics, while its DH methodology is about opening up scholarly expertise itself through shareable, transparent processes not locked into technically complex, pre-established, or large-scale research frameworks.

Project Research and Advocacy Goals

WE1S uses topic modeling to explore the idea of "the humanities" in public discourse. A complex concept of the kind that Peter de Bolla treated in his 2013 *The Architecture of Concepts* (his main example: "human rights"), "the humanities" as they are perceived are both tightly bunched in academic disciplines and broadly dispersed in extra-academic domains. Discussion focused on "the humanities crisis," "the decline in humanities majors," etc. creates flash points in the discourse. Yet the overall heat map of articles about the humanities, WE1S discovers, also extends into vast stretches of warm or cool discussion about humanities subjects intricately interwoven into the background of other domains of social life. Even articles so seemingly unremarkable (yet fully remarkable when we think about it) as an obituary or wedding announcement can mention the humanities as part of their *donnée*. WE1S seeks to open to view this whole conceptual architecture of "the humanities" as it exists in robust, living relation with culture at large.

Project Methodological Goals

Due to the lack of widely shared technical conventions and appropriate scholarly and publishing practices, today it is very difficult for a DH scholar to answer with documentation such questions as: *Where did you get those thousands of works in your corpus? Where did the metadata come from? What steps did you take to prepare and process the material? How many variations did you try? Where in the process was it critical for there to be "humans in the loop"?* The WE1S project addresses a growing need for ways to share and reproduce data-workflow in digital humanities research in order to make DH comparable to "open science" (see Bare, 2014). Indeed, data-intensive work in the sciences offers an especially good paradigm because of the degree to which it makes workflow and provenance management itself a thoughtful research field (i.e., research *about*, and not just tools for, managing workflow and provenance) (e.g., see Gil et al., 2007; and Garijo et al., 2012). The WE1S project is developing a technical framework that explores how the digital humanities can evolve similar, but also necessarily different, *humanities*-adapted standards of openness, shareability, and reproducibility. What amount of data and metadata, in what detail, at which processing stages, with what accompanying scripts, and so on, should be shared to support rich and persuasive scholarly discourse based on digital humanities research in the future? How will the criteria of "reproducibility, replication, and generalizability" (on the different shades of meaning of these terms in the sciences, see Bollen et al., 2015) join more traditional ideas of excellence in the humanities (e.g., "critical rigor") in the various contexts of collecting, curation, exhibition, editing, analysis, interpretation, and other work?

Overview

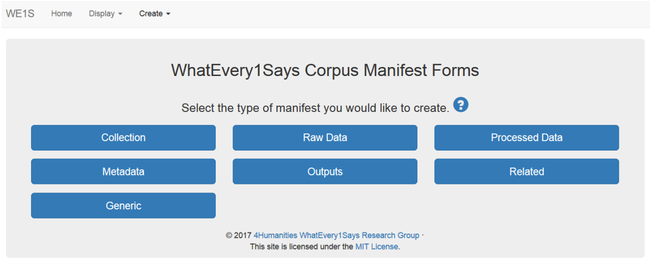
Alan Liu

WE1S explores public thought about the humanities, especially as mediated in journalistic articles that stage a dialogic relation between leaders in government, business, universities, the arts, and others with citizens. The project's end goal is to use such research to guide humanities advocacy. But rather than create a one-off project, the WE1S group has developed a robust technical and interpretive methodology comparable (though customized for DH) to scientific data workflow management systems (e.g., Apache Taverna, Kepler, Wings), provenance tracking systems (e.g., ProvONE), and similar schemes (see Gil et al., 2007; and Bose and Frew, 2005).

Manifest System for Data Provenance and Workflow Management

Scott Kleinman

The WE1S manifest schema uses a JSON Schema-based model to produce "manifests" that document the provenance of articles studied by WE1S as well as later transformations of the data, tying together scripts, stop word lists, outputs, visualizations, etc. used in project work. Manifests make the workflow transparent and facilitate on-the-fly reiterations or adjustments--e.g., staging a subset of the WE1S corpus for topic modeling or defining variant numbers of topics. Manifests are human-readable JSON files that are highly interoperable with other systems; they can be used programmatically to drive scripts or to crosswalk information to other workflow tools or metadata frameworks. The WE1S workflow management system uses the manifest schema to generate web forms, enabling non-technical users to create and query manifests, which are stored using the same JSON-like format in its MongoDB database. The system is easy to deploy and can be adapted for other humanities research projects simply by modifying the manifest schema.



WhatEvery1Says Corpus Manifest Forms

Select the type of manifest you would like to create.

© 2017 4Humanities WhatEvery1Says Research Group ·
This site is licensed under the MIT License.

Fig. 1: Web interface for WE1S manifest system.

Virtual Workspace Manager for Integrated, Containerized Data Manipulation and Processing

Jeremy Douglass

To address a range of computing demands from geographically distributed participants, the WE1S Workspace Manager facilitates open, reproducible DH research through a defined computing platform, a shareable online environment, integrated customizable workflows, and on-demand publishing of results. Tools for topic modeling workflows are configured on a virtual machine (a Docker container). Open data science notebooks (iPython / Jupyter) are the interface. Project templates are collections of notebooks (Python, R) chained into flows. Each new data exploration flow customizes a template, imports manifest data (from the WE1S manifest system), builds a topic model (Mallet), generates a visual browser (Andrew Goldstone's dfr-browser), publishes the browser to an interactive website, and packages a project for download and offline viewing. WE1S hosts a shared workspace online; it also runs on a laptop. Design and implementation of this virtualized, integrated workflow environment may be relevant to other DH projects, and is consonant with the philosophy of such other online or containerized integrated systems as Lexos or DH Box designed to make advanced DH research environments accessible.

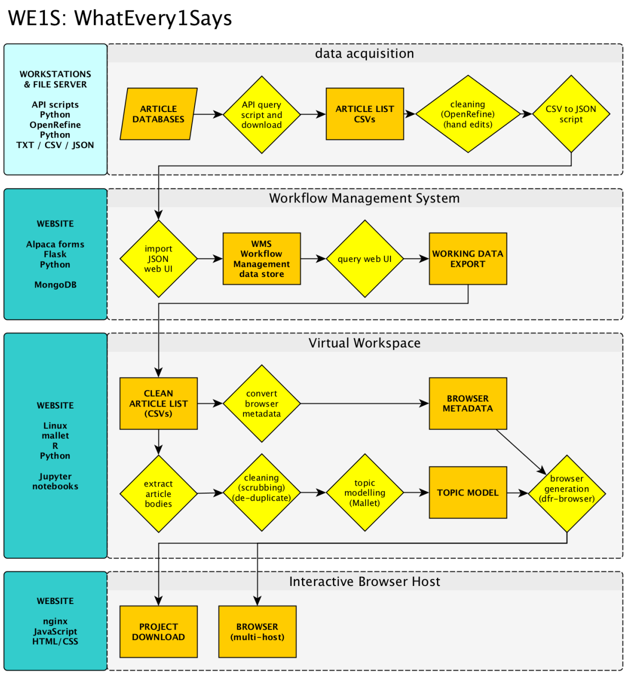


Fig. 2 Architecture of WE1S virtual workspace manager for integrated topic-modeling and visualization as implemented in a Docker virtual machine.

Constructing a "Random" Comparison Corpus

Lindsay Thomas

In public discourse, there are no natural boundaries between what does and does not count as "humanities-related" discussion. The humanities, for example, can appear in both precise and general ways: as a focal topic, as part of arts and culture, in particular forms (such as literature), as part of social and ethical concerns, as part of the biographies or obituaries of individuals, etc. Indeed, it may be that one feature of the humanities is their capacity to forge multiple links between tightly focused and general themes. There is thus no pre-definable "control corpus" of public discussion on the humanities that can serve as ground truth for WE1S's topic modeling experiments. WE1S is thus using a sampled, "random" corpus as a snapshot of the larger, unclosed set of media articles to assist in exploring by contrast what articles can sensibly be defined as "humanities-specific." Doing so not only constitutes a novel approach to topic modeling in the digital humanities; it also reveals intriguing issues about the philosophy behind statistical randomization (see Holland, 1986). This part of the panel discusses the "random" corpus WE1S created, its use within the WE1S project workflow, and includes theoretical reflections on incorporating methodology borrowed from the sciences and social sciences in DH work.

"Interpretation Protocol" for Topic Models.

Ashley Champagne

One of the needs in DH research is a for way to declare not just technical but *interpretive* workflows so that they can be shared, reproduced, and evolved by the research community. In the case of DH topic model studies, for instance, rarely are there transparent descriptions of the interpretive assumptions, steps, and iterations needed to decide how many topics to seek, what topics are interesting, how the topic model guides the human interpreter back to specific articles for examination (and vice versa), and how groups of researchers collaborate in using a topic model to generate hypotheses or come to conclusions. WE1S has created an initial declaration of its topic-model interpretation process that defines step-by-step interactions between machine learning and human interpretation/collaboration (e.g., when in the process humans convene to interpret a topic model; what outputs, visualizations, and secondary algorithmic products such as clusterings are used; how humans discuss a topic model; how topic models and interpretive acts are iterated; etc.). The goal is to make it possible for the larger DH community to improve or vary the topic-model interpretation process in open, shareable ways.

Prototyping How the WE1S Project Can Guide Humanities Advocacy

Jamal Russell

When WE1S has completed its topic models and interpretive studies, it will produce a public-facing site allowing others to explore the models and follow links to the original articles. But how can the project fulfill its ultimate ambition of guiding humanities advocacy? This final part of the panel reports on a unique early experiment in applying WE1S research. In 2016-17, a funded group of undergraduates studied sample articles from the WE1S corpus under the guidance of the project's topic models. They wrote a white paper on their findings with recommendations for humanities advocacy. And they created practical advocacy projects based on those recommendations. Using this concrete example as a springboard, the panel concludes by reflecting on the relationship between interpreting topic models and creating publicly accessible narratives about the humanities.

Bibliography

4Humanities: Advocating for the Humanities. Home page, n. d. http://4humanities.org.

4Humanities: Advocating for the Humanities. "'What Every One Says About the Humanities' Research Project (WhatEvery1Says)." 25 April 2013. http://4humanities.org/2013/04/what-everyone-says-about-the-humanities-research-project.

Apache Taverna (Taverna Workflow System). Home page, n. d. https://taverna.incubator.apache.org.

Bare, C. (2014). "Guide to Open Science." Digithead's Lab Notebook, 9 January 2014. http://digitheadslabnotebook.blogspot.co.uk/2014/01/guide-to-open-science.html.

de Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights.* New York: Fordham University Press.

Bollen, K., J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. I. Olds. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science -- Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences." National Science Foundation, 2015. http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Bose, R., and J. Frew. (2005). Lineage Retrieval for Scientific Data Processing: A Survey. *ACM Computing Surveys* 37.1: 1–28. https://pdfs.semanticscholar.org/3a05/2feb019328487068c8efc4c5dced8eb51a87.pdf.

DH Box. Home Page, n. d. CUNY Graduate Center. http://dhbox.org.

Garijo, D., P. Alper, K. Belhajjamey, O. Corcho, Y. Gil, and C. Goble. "Common Motifs in Scientific Workflows: An Empirical Analysis." 2012 IEEE 8th International Conference on E-Science, 2012: 1–8. DOI: 10.1109/eScience.2012.6404427.

Gil, Y., et al. (2007). "Examining the Challenges of Scientific Workflows." *IEEE Computer*, 40.12: 24-32. https://pdfs.semanticscholar.org/e45d/4aedd10229cbbeef8b2ec009f87ae1a4065e.pdf.

Goldstone, A. dfr-browser, v. v0.8a (June 8, 2016). Home page, n. d. http://agoldst.github.io/dfr-browser/.

Holland, P. W. "Statistics and Causal Inference." *Journal of the American Statistical Association*, 81.396 (1986): 945-960. http://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354.

iPython Notebook. (See Jupyter Notebook.)

Jupyter Notebook (formerly iPython Notebook). Home page, 5 March 2017. http://jupyter.org/.

Kepler Project. Home page, n. d. https://kepler-project.org.

Kleinman, S., LeBlanc, M.D., Drout, M. and Zhang, C. Lexos. v3.0. 2016. https://github.com/WheatonCS/Lexos/. doi:10.5281/zenodo.56751.

Lexos. (See Kleinman et al.)

Mallet. (See McCallum, A. K.)

McCallum, A. K. "MALLET: A Machine Learning for Language Toolkit." 2002/2016. http://mallet.cs.umass.edu/.

ProvONE. "ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance -- Unofficial draft." 27 March 2014. http://vcvcomputing.com/provone/provone.html.

WINGS (Semantic Workflow System). Home page, n. d. http://www.wings-workflows.org.

Computer-Assisted Conceptual Analysis of Textual Data as Applied to Philosophical Corpuses

Jean Guy Meunier
meunier.jg@gmail.com
Université du Québec à Montréal, Canada

Louis Chartrand
lochartrand@gmail.com
Université du Québec à Montréal, Canada

Mathieu Valette
mvalette@inalco.fr
National Institute for Oriental Languages & Civilizations
France

Jackie Chi Kit Cheung
jcheung@cs.mcgill.ca
McGill University, Canada

Marie-Noëlle Bayle
mrcal.bayle@gmail.com
Université du Québec à Montréal, Canada

Overview

Jean Guy Meunier and Louis Chartrand

Many DH projects call upon computer tools for descriptive and analytic purposes: lexicon statistics, concordances, parsers, descriptive statistics, classifications, topic modeling, annotations, automatic summarization, visualization tools, etc. They have been mainly applied to literary, political, journalistic, or otherwise mediatic corpuses. However, less work has been done on philosophical corpuses, or corpuses that have been tailored for the ends of philosophical investigation.

Computational approaches specialized for highly theoretical and abstract texts have proved their efficiency in various domains, particularly in those pertaining to the encoding, curation and presentation of textual data (e.g. digitization, web publishing, authorship analyses, etc.). These successes open avenues for more complex tools and methods to study linguistic features which are not directly observable, such as narratives, themes and concepts.

Concepts, in particular, constitute a key issue, given, on one hand, the polysemy of the concept of concept, and, on the other hand, its widespread use in philosophy and other disciplines. A conceptual analysis is a process through which we decompose and thus elucidate the meaning of a concept. While it is traditionally practiced from the armchair, concepts' meanings are reflected in the texts where they are expressed. As such, the development of methods and approaches to computer-assisted conceptual analysis of texts (CACAT) has the potential to make conceptual analysis more precise, more reliable, more exhaustive and more inclusive.

These new challenges call for an appropriation of modern computational tools which have demonstrated their potential at discovering such entities for natural language processing. The last two decades have seen important developments in computational linguistics and in machine learning which have enabled researchers to detect and manipulate various complex features of textual data. Complex objects such as entities, events, topics, arguments, or syntactic and discourse relations can now be detected and studied. Progress in modelling and learning approaches from fields like probabilistic modeling and neural networks have made it possible to represent complex representations between various latent and explicit textual features, and to learn them in efficient ways. Because they capture different aspects of concepts, including those which appear to be latent or implicit, the innovations could open new and exciting horizons for conceptual analysis.

In order to exploit this potential, digital humanists must participate in the conception of the aforementioned innovations. The humanities have problems and conceptual tools of their own, which differ from those of computer scientists, and, as such, ought to be enunciated and translated into tasks for algorithms to fulfill. On the other hand, computational approaches to conceptual analysis pose specific problems, which must be addressed as new methods are developed: indeterminacy of the nature of a concept and of the method of analysis, complexity in the relation between and among concepts, diversity of interpretations, incertitude in the evaluation schemes, limited set of computation tools to explore conceptual structures, shallowness of visualization tools, etc. These difficulties call for work from social scientists and humanists.

Our aim in this session is to give an overview of the challenges, avenues and opportunities that are shaping CACAT's development. Each paper plays a specific role within this session, so that information in one paper may serve as context for another. Jean-Guy Meunier's paper reviews the challenges facing CACAT, serving as context for the papers that follow. Mathieu Valette's paper (which will be moved to second place) expands on this topic. Drawing on the study of concept formation in philosophy of science, he offers a criticism on the concept of concept, proposes an alternative and draws the implications for the relationship between concepts and their concrete expressions in texts. In working to bridge the gap between the philosophical conceptual analysis in texts and modern techniques of computational linguistics, these two papers work from the conceptual analysis pole. The following two papers, by Louis Chartrand and Jackie Cheung, work from the computational linguistic pole, presenting models and techniques which have the potential of addressing the challenges of CACAT. The final paper, by Marie-Noëlle Bayle, is a recent application of CACAT that exemplifies its potential in discovering implicit dimensions to a concept.

Modelling Computer assisted conceptual analysis in text (CACAT)

Jean-Guy Meunier

Conceptual analysis paradigms

In many fields of scientific research, be they social sciences, natural sciences or even professional practices, abstract or highly theoretical concepts are explored to discover their content and deepen the knowledge they embed. However, there is no consensus on the nature of a concept or on the methodology to analyze them. For example, how would one proceed in analyzing the concept of Evolution in Darwin's writings? Three radically different paradigms parameterize the methodology: philosophical, linguistic and cognitive.

In the philosophical paradigm, concepts are identified to the meaning of predicative words. For some, their analysis aims at finding the conditions (necessary, sufficient, fuzzy, etc.) under which these words refer to objects, events or actions in a possible or actual world. For others, analysis consists mainly of identifying the sense or intention of these words as related to the epistemic or metaphysical conditions for their understanding. Finally, for some others, an analysis should consider the use and context (linguistic, social or other) of these words. Hence, in this philosophical paradigm, conceptual analysis becomes a sort of logicopragmatic analysis of the meaning of words. In our Darwin example, this paradigm would therefore ask what are the meaning conditions of the word evolution when Darwin uses it.

In the linguistic paradigm, concepts are also related to the meaning of words. For the Saussurian structuralists a concept is the core meaning embedded in the structure of the signified (le signifié) of words. For the neo-structuralists, the generativists and the cognitivists, a concept is also equated to the semantic content of predicative linguistic expression, and meaning is understood as a complex set of semantic properties (features, relations, frames, nets, etc.,) underlying isolated words or their position in sentences and discourse. Here, conceptual analysis becomes identified with classical semantic analysis of words. In Darwin's works, the analysis would explore the semantics properties of the English word evolution: for instance, it would study its lexical content, its synonyms, its topics, is semantic nets, etc.

In the cognitive paradigm, concepts are the results of cognitive or mental operations. For psychology, they are seen as a sort of cognitive categorization. For the analytical and hermeneutic traditions of philosophy of mind, they are mental states or world representations. Conceptual analysis consists then in exploring how semiotic or linguistic forms embed categories, intentions, conceptual spaces, beliefs, mental states, Weltanschauung, etc. Hence conceptual analysis bears resemblance to an exploration of cognitive operations or states: representing, categorizing, reasoning, argumenting, entailing, etc. In our analysis of Darwin, this cognitive paradigm would focus the analysis on the mental operations underlying the meaning of evolution. How is this category of mental representation acquired, built reasoned on, argued, etc.?

Choosing a paradigmatic methodology for analyzing concept is difficult, then, because not one of them is canonical. Conceptual analysis becomes an even more acute problem when computations are introduced in the methodology. The level of complexity of the task is so high that is not obvious how a computer assisted conceptual analysis of text (CACAT) project can be realized. Should it be computer-tool-driven or model-driven?

Tool driven approaches

The first type of approach is tool-driven. Once a methodology inspired by one of the paradigms presented above is chosen, its practitioners use some computer programs already built and inserts them in appropriate moments of the analysis procedure. Many computer tools for this task actually exist.

A first set of tools focuses on the lexical expressions of a concept. The most classical ones are concordancers, collocation and lexical analysers, taggers, etc. These tools explore the lexical properties and contexts of one or a few canonical predicates, expressing the specific concept to be analyzed. The limits of these types of tools lie in their underlying design hypothesis: a conceptual content is to be explored through specific canonical expressions. Such a hypothesis restricts the exploration of the conceptual content to one or a specific number of predicates. This is problematic, for as we know, concepts can be expressed in language in a myriad of ways. For example, it would be very problematic to restrict Darwin's concept of evolution to the analysis of the word evolution alone. Secondly, they may produce results that are larger than the original text. This is the case of the concordance of the concept of Esse in the Thomas Indexicus. Finally, sometimes, the opposite happens. These tools may deliver only a fraction of the overall textual segments or word collocations whose content is pertinent. For instance, in Darwin only uses a few dozen times the lexical form evolution. Hence concordance, collocation, etc. on such a small sample are not very fruitful for a conceptual analysis.

A second set of tools highly influenced by classical AI approaches focuses on natural language processing (NLP). These tools are sensitive to various meaning aspects of words, such as their semantic definition, their encyclopedic, pragmatic discursive content, etc. They promise to deliver finer results for a conceptual analysis. But these tools also have limits. Their underlying hypothesis is that these semantic, pragmatic and encyclopedic information added in the grammar and the lexicon will enhance the exploration of conceptual content. Unfortunately, the added information has often been collected from common and ordinary semantic knowledge of shared language usages. Such tools will then often tend to identify already known properties belonging to this common information about the lexical conceptual word under inquiry. And most of the time, it will ignore the properties that precisely are the one that are specific to the concepts analyzed mainly when they are original, and belong to a reflexive, creative literary or reflexive discursive process, etc. These semantic properties would not be part of the common doxastic conceptual content. For instance, a philosophy scholar using such types of tools would not be very satisfied in discovering that Darwin's concept of evolution is a name meaning an action of the type change and applied to the object: natural species.

Recently, a last set of tools that are more mathematically grounded, such as neural net and Bayesian classification, vector semantics, machine learning, deep learning, etc., have become appealing and are used in language processing, They can process large data and learn semantic information by themselves. But like the other set of tools they have their limits. First, they are nor readily usable. They are in fact very complex algorithms, and are not easily mastered by humanities scholars. Secondly, their lack of traceability becomes a major obstacle when applied to large and theoretical textual data where results become difficult to evaluate. Thirdly, they seem more successful for information retrieval applications than for digging into deep conceptual content. For the moment, we are not sure that how they can effectively assist conceptual analysis.

From these remarks, it does not seem to us that conceptual analysis can only be a serendipity tool-driven approach. The results produced by these tools have not yet convinced the scholarly community that practices expert conceptual analysis.

Model driven approaches

The second type of approach is model-driven. Recent philosophers of science such as Morgan and Morrison (1999), Giere, (1999), Leonelli (2007) for instance, see science as a building models process where models are heuristic means for describing, explaining and understanding reality. And Mc Carthy (1999), has seen this modelling approach as a means of better understanding digital humanities interpretative projects. For our part, we explore this hypothesis and see CACAT as type of scientific inquiry where various models are used as intermediaries for understanding the analysis of highly theoretical and abstract concepts. In this perspective, we distinguish four types of models: conceptual, formal, computational and experimental.

A conceptual model defines parameters for identifying, explaining and understanding the properties and structure of linguistic items expressing conceptual content. A formal model translates certain aspects of a conceptual model in some controlled formal language that describes or identifies properties and relations of these conceptual expressions. A computational model translates some formal expressions of the formal model into algorithms and programs. Finally, an experimental model designs implementation of these formal models in a concrete computer where the analysis can be simulated and ultimately evaluated in correspondence to the other models.

In a concrete procedure, all these models interact and can be modified and adjusted. This allows the inquiry to be controllable and repeatable. It has been our own experience that, if a computer assisted conceptual analysis project is to be successful it must construct at least these four models. A CACAT project cannot bypass these models and their interactions.

Designing these models, their interactions and their experimentation to see CACAT as a scientific endeavour and not just computer gadget exploration. But each model is not built easily. And nothing comes smoothly. They are part of the research process. And much work must be done to clarify the conceptual, formal, computational and experimental models pertinent for a successful and pertinent conceptual analysis.

Digital epistemology for concept analysis

Mathieu Valette

In the humanities, theory is most of the time outlined with texts: papers, books, conference presentations, lectures etc. we claim that the scientist is first a reader and a text producer. This textuality is so ordinary that it is almost invisible, and, as such, not considered as an object of science. Moreover, theories are read as synchronic systems, or even achronic systems, depending on their specific purposes (describing one fact, explaining one phenomenon...). Scientists appropriate models and concepts like tools; they have to know their function and how to manipulate them, but they do not care about knowing practical details of their enunciation. In fact, they ignore them, more or less. They find such details embarrassing, because they make concept borders fuzzy: lexicons, glossaries, and also handbooks, as they extract the concepts from their context, and standardise the definitions, creating an illusion of stability and tangibility. But concept textuality necessarily has an incidence, not only on interpretation, but also on theorisation. If the scientist is a text producer, then theorisation is the construction of meaning. Theorisation is forced by enunciation, and scientific works, beyond their materiality, can be considered as text.

The textual aspect of scientific works had been noticed by those in Europe looking at epistemological culture. In this respect, French philosopher Michel Foucault's works, in the 1960s, must be acknowledged (see e.g. Foucault 1969). Foucault put in place a philological analysis of discourse centred on the combination and evolution on specific discursive structures. His purpose is, firstly, to recognize "discursive formations", i.e. stabilized relations, regularities between objects, types of speech act, concepts and topics; and, secondly, to recognise breakpoints in idea system history. Foucault followed the example of some of his famous predecessors, such as Gaston Bachelard, Georges Canguilhem and Martial Gueroult. Bachelard's notion of Epistemological break, or Canguilhem's notion of concept shifts shows, for instance, that the history of a concept is not that of its increasing rationality and refinement, but that of the different fields in which they have been designed and validated. What we will call digital epistemology is a linguistic approach to this style of French epistemology.

Our topic is the study of scientific texts using, on the one hand, corpus linguistics tools which have been developed over the 40 last years and, on the other hand, a linguistic methodology (see Rastier 2009, Valette 2003). Thus, our purpose is to develop tools and methodology Foucault did not have, among other reasons because some textual phenomena—as, for example, lexicon evolution, which depends on the reader's subjectivity—are invisible to a classical philological analysis. Concept emergence, concepts' individual and inter-related evolutions, the appropriation of a specific thematic, palinode, etc. constitute further examples. We do not adopt the logician's position, considering that conceptualization is a linguistic phenomenon with its own construction rules linked to a particular function of language. Neither do we ignore the psychological, social and interactional reasons of the development of concepts. Firstly, we consider that textuality—i.e. the constraints of the textual layout, formulations, be they constraints of syntactic, semantic, lexical or related discursive traditions (including genres and speech)—plays a major role in concept formation. Secondly, we do not consider texts only as resources to mine and extract terminological and conceptual material, but as archives, or, in other word, as the objective tracks of the process of creating concepts.

In essence, we focus here on concept emergence considered as the result of a slow and gradual stabilisation of contextual semantic feature. Drawing on recent critical readings of Saussure's semiology (see Rastier 2015), we propose to consider a concept as a stabilized semantic form; that is, as a combination of semantic features (or semes) mainly inherited from various contexts in which it has occurred. Eventually, we link concept design with text production rather than identification of items in a general ontology (Valette 2010).

Topic models for conceptual analysis

Louis Chartrand

The last two decades have seen the rise of topic models in natural language processing (NLP). From the early successes of Latent Semantic Analysis, which decomposes datasets into "conceptual" dimensions, the introduction of probabilistic and generative models have enabled the discovery of underlying structures that condition the lexicon of a text. Those structures, in turn, are used to construct meaningful representations of corpuses and documents, and have proven fruitful in improving performances in many NLP tasks.

Those tools have interesting potential for the Digital Humanities, as they discover entities which are, on one hand, robust features of textual data, and, on the other hand, easily representable and interpretable by humans. For instance, topics may help in tasks such as categorizing documents or selecting a relevant sub-corpus for analysis. However, once topics are represented using the words to which they are likely to be associated, they can also be used to make sense of what a set of textual segments are about, or to visualize the evolution of discourse in a corpus through time. As such, topics have interesting potential when it comes to representing textual data and improving our analyses of it.

In this presentation, some prominent topics models—LSA, LDA and DTM—will be presented and contrasted, and their potential uses for Digital Humanities will be discussed.

Latent Semantic Analysis (LSA)

Introduced by Deerwester et al. (1990), LSA used tools from linear algebra, in particular singular vector decomposition, to transform the representation of text segments in the form of word counts to a representation in the form of participation to "concepts", or semantic dimensions.

As words give us a good idea of what a text is about, it is common practice in text mining to represent text segments by counting its words. A document containing "apple", "orange" and "pear" a high number of times each is likely to talk about fruits. And if multiple documents share the same words, they are likely to share common topics. However, this approach does not fare well with synonyms, which it does not recognize.

The LSA uses co-occurrence in word uses to synthesize the word-count representations into more compact semantic dimensions. As synonyms tend to have the same cooccurents, they also tend to participate to the same semantic dimensions. As a result, the new representation is closer to a semantic representation than was the word-count representation, hence the name "Latent Semantic Analysis".

Latent Dirichlet Allocation (LDA)

While LSA still is part of every NLP reputable toolset, it falls short in at least two key aspects. Firstly, its semantic dimensions are hard to read for a human: from a list of its most prominent words, it is usually hard to give a satisfying interpretation of what a semantic dimension is about (Chang et al. 2009). Secondly, it has no clear hypothesis as to how text is structured. On one hand, this makes it harder to explain semantic dimensions in terms of linguistics, psychology or discourse analysis. On the other, it means that LSA gets only part of the picture, and better algorithms with additional assumptions might produce better semantic dimensions.

Latent Dirichlet Allocation (Blei et al. 2003) is a probabilistic models which attempts to address this latter issue, and ends up addressing the former as well. It supposes that in a corpus, there is a certain number of topics, which, when activated, make it more or less likely for specific words to be present. Thus, when someone writes a document, LDA assumes that she selects a certain restricted number of topics, which in term condition which words will be found in the document. Using this assumption and an arbitrary number of topics, the algorithm infers the most likely list of topics, and their most likely assignments to documents.

As such, it produces once again a representation of documents or text segments from word counts, but in terms of topics rather than more abstract conceptual dimensions. The words most likely to be present when a given topic is activated are often visibly related, either semantically or because they participate in a transparent narrative. As such, they are easily read by a human interpreter, and can be used to give a sense of what documents, sub-corpuses or textual segments are about.

Dynamic Topic Models (DTM)

Another boon of LDA is that its probabilistic model can be modified account for particularities of the corpus, or to model features that we want to study in particular. For example, if we have a corpus that spans across decades, we might expect topics to evolve with time, as society and culture change.

To model this, the algorithm devised by Blei & Lafferty (2006) uses a corpus split in time slices (say, per year) and topics are split accordingly, such that topic 1 at time 1 is different from topic 1 at time 2. Then, a Markov assumption is enacted on the time series: topic 1 at time 1 conditions topic 1 at time 2, which conditions topic 1 at time 3, etc. This gives topics the freedom to evolve, while enforcing a certain degree of conservatism.

Using this, one can not only track topics more efficiently, but also see the evolution of topics across time.

What is a topic?

What is it, however, that we are talking about when we speak of LSA's conceptual dimensions or LDA's topics? Can it be equated with the notion of TOPIC that we encounter in discourse analysis, for instance?

While there are a variety of definitions for words such as "topic" and "theme", most agree that a topic is what a text is about (Rimmon-Kenan, 1995). On this score, LDA's topic does seem to agree with the common notion of TOPIC: a word list representing a LDA topic is read as a representation of what the textual data is about (Blei, 2013). Furthermore, the information the probabilistic model captures is the one that is redundant across a number of text segments. As such, it highlights words and concepts which keep coming back as they put in relation with various entities in sentences. In other words, textual discourse and narratives are being sewn around them.

On the other hand, humans tend to make slightly different representations of topics compared to machines (Chang 2010), more readily constructing topics around concepts and thus providing sparser (more compact) representations. As Chang suggests, this might be because humans build these representations using general domain knowledge, whereas topic models try to infer this knowledge from word distributions. This seems to tell us that we should understand LDA topics as indicators, or reconstructed traces, of the topics that underlie a text, but not as true representation of topics themselves.

Using topic models in conceptual analysis

As Chang's 2010 experiment suggest, topics entertain special relation with concepts, as a topic tends to be associated with a restricted number of concepts which are expressed very often in the text.

As such, topic models' potential in representing textual data can be exploited to discover associations that are likely to be useful to conceptual analysis and other philosophical analyses. For instance, it can help the analyst identify the most important parts of a corpus, and those that can be discarded. They can also be leveraged to build representations of the contexts in which the concept of interest appears, thus giving a sense of the topics with which it is associated. Using DTM, one can also get a sense of the evolution of a concept within a diachronic corpus. Beyond discovery of new informations concerning a concept's expression in a corpus, topic models can be useful to test some association, as the structures they uncover are relatively robust.

That said, as the DTM model shows, topic models can be used in a large variety of use cases, as their model can be expanded to take into account a corpus' metadata and thus open new and innovative avenues for conceptual analysis and the Digital Humanities in general.

Unsupervised natural language processing for conceptual analysis of events

Jackie Chi Kit Cheung

In unsupervised machine learning, an algorithm is trained to discover regularities in data without access to human-provided labels. Such techniques can be useful in conceptual analysis of text, in cases where we do not have or want to impose a schema on the text corpus under analysis. The basic intuition behind unsupervised natural language processing techniques is that objects that appear in similar contexts in the data should be assigned similar representations, such that they can be grouped into clusters.

Unsupervised models differ according to several characteristics, from the type of information that is made available to the learner, to how similarity is defined between the different objects that are modeled, to the expected form of the output cluster that is learned. For example, the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) is a probabilistic model which is given access to multiple documents for training. The crucial assumptions behind the LDA model are that each document can be described as a mixture of multiple topics, and each word in a document is generated by one of the topics in that mixture. As a result of training an LDA model, multiple topics are learned, which correspond to clusters of words that tend to co-occur in the same documents.

More recently, there have been a number of unsupervised models that have been used to discover the structure of a sequence of entities and events that appear according to some narrative in the natural language processing literature (Chambers and Jurafsky, 2008; Cheung and Penn, 2013). This is accomplished by explicitly modelling the sequential dependencies of events as they appear in a document. I will provide an overview of the assumptions of the event structure being learned by such models. For example, some methods produce discrete sequences of prototypical event and participant roles. In the work of Chambers and Jurafsky, (2008), narrative chains are learned corresponding to prototypical roles in a narrative. A chain such as _ accused X; X claimed _; X argued _; _ dismissed X might correspond to a defendant in a trial. Other work frame the problem as a task for probabilistic learning. Cheung and Penn (2013) define a probabilistic sequence model, in which the structure of an event and its participants are explicitly represented in the model as latent random variables. The nature of a learned cluster, then, would be how it influences the conditional probabilities of generating other cluster labels, as well as the word emission distributions from that latent topic (as in an LDA model).

I will discuss how such models can be used to discover templates of prototypical events, including how events and event participants are typically expressed in language. Such approaches can easily be applied to multiple domains, including texts in the legal or medical genres, because they make minimal assumptions about the structure of events, and do not require training data. I also discuss other applications of these models to information ordering, and automatic summarization, which may be of interest to researchers in conceptual analysis for the digital humanities.

A computer-assisted analysis of SYMPTOM in psychiatry

Marie-Noëlle Bayle

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) is a general classification and diagnostic tool used in the rich and diverse universe of mental health. Being widely distributed and available online, it allows everyone to have a direct access on how to make a psychiatric diagnosis. To facilitate its reading, laypeople and professionals alike may consult definitions for important notions in the glossary section. However, these few lines will often fail to capture the complexity of a term. For instance, at the core of the clinical assessment of a disorder lie its signs and symptoms. Therefore, a proper understanding of what a symptom means, and how this concept relates to the disorder, is essential to the diagnostic approach.

In the DSM-5 glossary, symptom is defined as "A subjective manifestation of a pathological condition." In the practice, it is often treated as a necessary and/or sufficient condition for to seek diagnosis, or as a constraint on possible diagnoses. As such, non-doctors and patients often think of a symptom as a singular event, and conceive of it purely in terms of content and a means to a diagnosis.

However, as experience of mental disorders is often messier, we may wonder if such a notion of SYMPTOM overlooks important features of the way this notion is actually reflected in the DSM-5. Our hypothesis is that both the glossary definition for this concept and the common understanding of this notion fail to account for such essential aspects. To provide evidence for this claim, we performed a computer-assisted conceptual analysis of text (CACAT) for the concept SYMPTOM in the DSM-5. We find evidence that SYMPTOM is strongly associated to a dimension of temporality, and that it is expressed in relation not only with the disorder, but also remission. As such, it is proposed an improved definition would not only better reflect the content of the DSM-5, but could also contribute in a better understanding of assessment and practice of diagnostic.

Method

The dataset consists in a small corpus composed with the most relevant chapters of DSM-5. Computational Text mining method and manual qualitative approach were used in a processing chain for the conceptual analysis of SYMPTOM. Firstly, extraction of the textual data from noisy sources and cleaning of the text were performed. Secondly, all sentences with the term "symptom" in it, plus one sentence before and after, were extracted, yielding a set of textual segments, on which stemming and lemmatisation were performed. Thirdly, the pretreated data was used to create a document-term matrix with the TF-IDF weighting scheme. Fourthly, from the matrix, textual segment clusters were produced with the k-means algorithm. Fifthly, the most salient words in each cluster and in the whole subcorpus were first represented using word clouds. Finally, relations of similarity between the most relevant words in each cluster and in the subcorpus were represented in a 3d space. All steps were performed using common R modules (tm, RWeka, qdap, cluster, knn, ggplot2, rgl). Each cluster was interpreted as a specific field in which a hypothetical conceptual property of SYMPTOM is expressed. Categorization was done by annotating manually the most typical textual segments in every cluster according to cosine distance to the centroid. Annotations consisted in the main conceptual property of SYMPTOM expressed in a segment. Syntheses of these annotations were done for each cluster.

Experimentation

From the subcorpus, 2036 sentences containing the term "symptom" were extracted which contained 5761 different word types. The words most associated with symptom in this subset of the corpus are disord (disorder), criteria, presen (presence), sever (severity), medic (medical, medication). Using k-means, 30 clusters were extracted but 17 are deemed noisy, as they contain 10 textual segments or less. Most of the remaining clusters are located in the section II of the DSM (diagnostic criteria and codes), several having most of their segments in a specific disorder chapter. The converse, however, does not hold.

Discussion

Analysing those clusters reveals that SYMPTOM has a transdiagnostic property, and is not only defined by its specific content. For example, let us examine cluster #8, which contains 98 segments, 90% of which fall in the three chapters about psychotic and mood disorders. Symptom in this cluster is linked with the words depress, hypomania, mania, but also with episode, period, full, meet. Furthermore, annotation of joined documents shows that temporality is a conceptual property of SYMPTOM. It modifies the pathological dimension of its content. SYMPTOM is not in direct causal relation with DISORDER; it is a dynamic sign whose presence needs to be situated in an episode, regardless of whether its content is depressive or manic. Therefore, the mere presence of a symptom is not sufficient for a diagnosis. Conversely, SYMPTOM is also linked with the concept of (partial) remission. SYMPTOM and DISORDER have a complex relationship on a continuum between negative (remission) and positive (disease) poles.

In conclusion, a mixed method, combining computational and manual processing and using quantitative and qualitative approaches, was applied in our conceptual analysis of the concept SYMPTOM. SYMPTOM appears to be a more complex and dynamic concept than patients and other non-doctors usually understand it to be. As a result, a better understanding of this complexity would likely profit assessment, diagnosis and treatment of mental disorders

Bibliography

American Psychiatric Association, (2013). Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC, American Psychiatric Association.

Bachelard, G. (1938) La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective. Vrin, Paris; The Formation of the Scientific Mind. Clinamen, Bolton, 2002

Blei, D. M. (2012) Probabilistic Topic Models. Communications of the ACM 55, 4:77.

Blei, D. M. (2013) “Topic Modeling and Digital Humanities.” Journal of Digital Humanities, April 8, 2013

Blei, D. M., and Lafferty, J. D. (2006) Dynamic Topic Models. In Proceedings of the 23rd International Conference on Machine Learning, 113–120. ACM

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research 3: 993–1022

Chambers, N., and Jurafsky, D. (2008) Unsupervised Learning of Narrative Event Chains. ACL, pp. 789-797.

Cheung, J. C. K., and Penn, G. (2013) Probabilistic Domain Modelling With Contextualized Distributional Semantic Vectors. ACL, pp. 392-491.

Chang, J. (2010) Not-so-Latent Dirichlet Allocation: Collapsed Gibbs Sampling Using Human Judgments.” In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 131–138. CSLDAMT ’10. Association for Computational Linguistics, Stroudsburg, PA, USA.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L. and Blei, D.M. (2009). “Reading Tea Leaves: How Humans Interpret Topic Models.” In Advances in Neural Information Processing Systems, pp. 288–296

Chartier, J. F., and Meunier, J. G. (2011). Text Mining Methods for Social Representation Analysis in Large Corpora. Papers on Social Representations, 20: 37.1-37.46.

Danis J. and Meunier, J. G. (2012). CARCAT : Computer-Assisted Reading and Conceptual Analysis of Texts : An experiment applied to the concept of evolution in the work of Henri Bergson. Digital Studies/Le champ numérique, 3(1).

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R.A. (1990) Indexing by Latent Semantic Analysis. JASIS 41, 6: 391–407

Foucault, M. (1969) L’archéologie du savoir. Gallimard, Paris; The Archaeology of Knowledge (1969), Routledge, 1972

Giere, R., (1999), Using Models to Represent Reality, in: L. Magnani, N. Nersessian and P. Thagard, (eds.), Model-Based Reasoning in Scientific Discovery. Plenum Publishers, New York, pp. 41–57

Hofmann, T. (1999) Probabilistic Latent Semantic Analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 289–296. Morgan Kaufmann Publishers Inc

Leonelli, S. (2007) What is in a Model? Combining Theoretical and Material Models to Develop Intelligible, Modeling Biology: Structures, Behaviors, Evolution. In Manfred Dietrich Laubichler, Gerd B. Mülle (eds.) Modeling Biology. MIT Press

McCarthy, W. (2004). Humanities Computing. Palgrave MacMillan

Morgan, M.S., and Morrison, M. (eds.), (1999), Models as mediators. Perspectives on natural and social science. Cambridge University Press

Rastier, F. (2001) Arts et sciences du texte. PUF, Paris.

Rastier, F. (2015) Saussure au futur. Les belles lettres, Paris.

Rockwell, G. (2003). What is Text Analysis, Really? Literary and Linguistic Computing, 18(2).

Salton, G., and McGill, M. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Turenne, N. (2016). Analyse de données textuelles sous R. ISTE Éditions, coll. Sciences Cognitives. Londres.

Valette, M. (2003) Conceptualisation and Evolution of Concepts. The example of French Linguist Gustave Guillaume, Academic discourse – multidisciplinary approaches, Kj. Fløttum & F. Rastier, eds., Novus Press, Oslo, pp. 55-74.

Valette, M. (2010) Des textes au concept. Propositions pour une approche textuelle de la conceptualisation », Actes des 21es Journées francophones d’Ingénierie des Connaissances (IC'2010) (8-11 juin 2010), Nîmes Sylvie Despres, éd., Publication de l’Ecole des Mines d’Alès, pp. 5-16.

Widdows, D. (2004). Geometry and Meaning. CSLI Publications, Center for the Study of Language and Information, Leland Stanford Junior University

Black Spatial Humanities: Theories, Methods, and Praxis in Digital Humanities (A Follow-up NEH ODH Summer Institute Panel)

Angel David Nieves
anieves@hamilton.edu
Hamilton College, United States of America

Kim Gallon
kgallon@purdue.edu
Purdue University, United States of America

David J. Kim
djkim@udel.edu
University of Delaware, United States of America

Scott Nesbit
snesbit@uga.edu
University of Georgia-Athens, United States of America

Bryan Carter
bryancarter@email.arizona.edu
University of Arizona-Tucson, United States of America

Jessica Johnson
jmjohnso@gmail.com
Johns Hopkins University, United States of America

Introductory / Framing Remarks

Dr. David J. Kim

Dr. Kim Gallon’s essay, “Making a Case for the Black Digital Humanities,” in the most recent Debates series volume argues for the continued development of a “relationship between the digital humanities and Africana/African American/Black studies … so as to highlight how technology, employed in this underexamined context, can further expose humanity as a racialized social construction” (Gallon 2016: 42). With the recent proliferation of projects including Black Gotham, Slave Revolt in Jamaica, 1760-1761, Freedom on the Move, and T-

RACES an emerging focus on the role that geospatial technologies can have in engaging with the history of race across the African Diaspora has become an important area of digital scholarship. This hybrid 90-minute panel is a timely follow-up to a summer 2016 NEH ODH Institute for Advanced Topics in the Digital Humanities, “Space and Place in Africana/Black Studies: An Institute on Spatial Humanities Theories, Methods and Practice for Africana Studies”. The panel includes Institute faculty in a broad discussion regarding rigorous, community-based, and applied methodologies that now constitute a focus on race in the spatial humanities. Panelists will limit their presentation to no more than 10 minutes followed by a 30 minute Q&A working session to discuss the broad parameters of the sub-field and future scholarly approaches.

Spatial humanities has transformed the work of researchers, enabling profound considerations of space in relationship to human behavior and culture across time and place. Art history, literature, history, philosophy and religion – all notable fields from across the humanities – have benefited from scientific and quantitatively oriented technologies and tools to better understand the intersections between space and the human condition. It is particularly timely now to question space in relation to African-descended people’s ability to traverse and negotiate spaces in western societies. The history of the Black body and American public and private space is particularly problematical as the presence of the Black body there has been largely unwanted during long stretches of American history. Indeed, both geographical and social spatial differentiation in the United States and the larger western world has largely been predicated on racial difference, exclusion and segregation. Leading proponents of this emerging sub-field, Gallon and Nieves, are slated to co-edit a new digital book series with the University of Georgia Press. Black spatial humanities is a sub-field of the spatial humanities that – in light of the fluidity of space in relation to people of African descent in the United States, Africa and the African Diaspora – studies the children of the Diaspora across specific places and times within an epistemological framework that is predicated on an ethic of anti-racism and liberation.

The panel will highlight some of the most innovative scholarship in both analog and digital formats that rely, in part, on new technology and forms of digital scholarly publication.

The panel will address the following questions:

- How do the spatial humanities and Africana/Black Studies work together to posit and practice a different way of knowing and imagining the world?
- How do racial identities impose a certain framework on our understanding of space?
- How can the spatial humanities help us experience the lived realities of Black/Brown bodies?
- How do maps/mapping technologies reflect and/or counter the realities (the dynamism) of Black life?
- How can spatial technologies provide us with a way of understanding the forms of inhumanity attributed to or placed upon Black/Brown people?

Modeling the Nineteenth-Century Colored Conventions Movement

Dr. David J. Kim

Throughout the nineteenth century, free and fugitive blacks across the US organized more than 150 state and national Colored Conventions. Less prominent than the contemporaneous abolitionist movement in US History writ large, the Colored Conventions movement represents a complex network of black political and religious leaders, intellectuals and entrepreneurs dedicated to the cause of racial justice in education, labor, citizenship and international human rights. As an introductory case study for the panel, this presentation will discuss the Colored Conventions Project (P. Gabrielle Foreman, faculty director) and its various digital and collaborative layers of archives building. Reflecting on its initial stages, as well as looking forward to the next phase of development, it will discuss the challenges of "modeling," as a digital humanities' methodological framing, both the interpretative and the community-building possibilities of this archive: from the discovery of documents to the design of the forthcoming database.

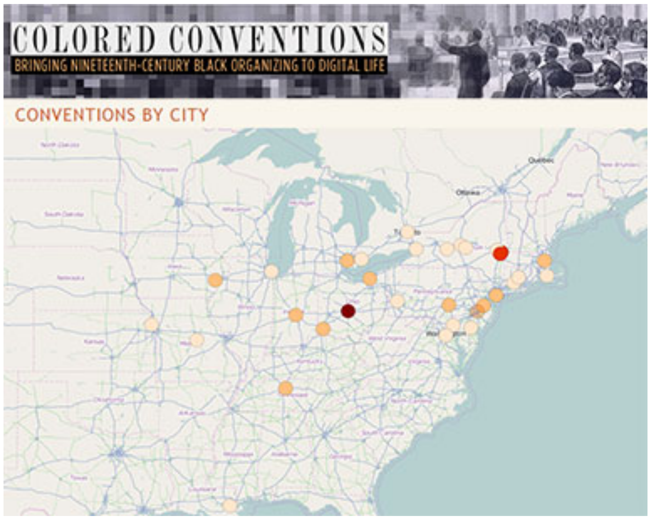

COLORED CONVENTIONS
BRINGING NINETEENTH-CENTURY BLACK ORGANIZING TO DIGITAL LIFE
CONVENTIONS BY CITY

Figure 1. Sample Map, Conventions by City, Colored Conventions Project

Virtual Harlem 4.0: Experiencing the Humanities through Virtual and Augmented Realities

Dr. Bryan Carter

As technology continues to evolve at a blistering pace, digital humanists strive to incorporate ways in which these new tools for research, data visualization, haptics and mobile technologies fit into our own work. Of the many exciting tech developments, augmented reality and virtual reality are poised to make a significant impact on the way we teach, research and experience the humanities. This talk will focus on new developments within the Virtual Harlem Project to include motion and facial capture, data visualization and four-dimensional (4D) learning experiences.

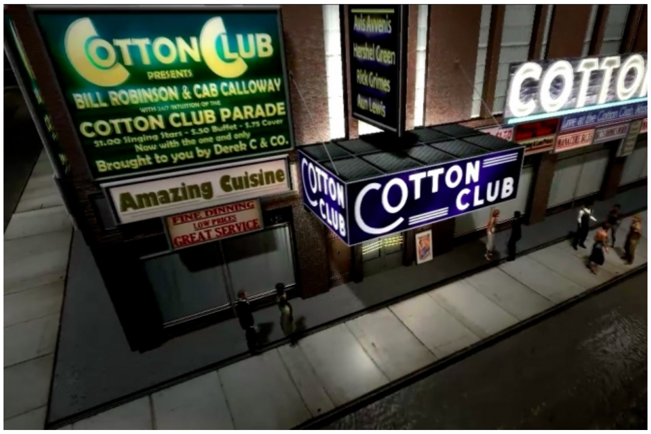


Figure 2. 3D Model, Cotton Club, Virtual Harlem 4.0

The Spatial and Digital Dimensions of the History of the Black Press

Dr. Kim Gallon

In geopolitical terms, Africana Studies scholars have studied everyday mobilities, which include the flow of people, objects and ideas backward and forward across the Atlantic Ocean. Black people's ability to move across time and space is a cornerstone for understanding their social condition. Thus, numerous scholars have documented and theorized the integral connections between black mobility and citizenship, freedom and resistance. This paper moves beyond geospatial dimensions to critically assess and examine literal spaces occupied by ideas and identities in the black press. At the same time, the use of digitized newspapers shows how black news as space can be mapped onto broader concepts concerning black people's struggle to make their humanity manifest, as well as to think critically about the various digital topographies of historical black newspapers.

Enslaved Women's Narratives in Eastern Virginia: Intersectional Approaches in GIS

Dr. Scott Nesbit

Much work in the spatial humanities depends upon the abstraction and reduction of historical processes, events, and, most problematically, actors, into data points, perhaps visualized as simple dots on a map. Scholars including Johanna Drucker and Miriam Posner have argued for more nuanced, humanities-based approaches to data visualization. Yet, if we reject inherited data visualization tools and statistics-based approaches, scholars may find at least some helpful analytical possibilities foreclosed. This talk will explore the tension at the heart of humanities data visualization in a black studies context by examining one particular question within the history of slavery and emancipation: who was escaping into United States lines during the American Civil War? The paper will argue for GIS-based approaches that are sensitive to the intersectional identities and the actions of formerly enslaved women in eastern Virginia by examining runaway slave advertisements.

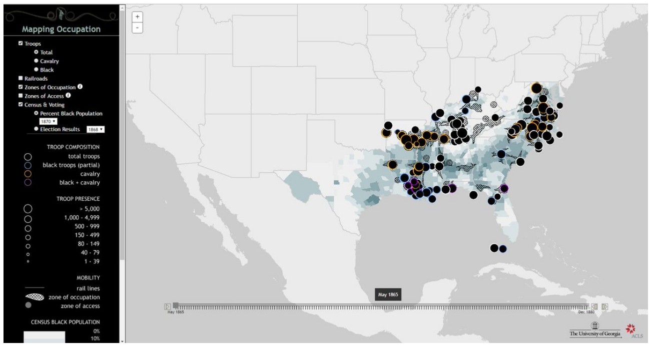
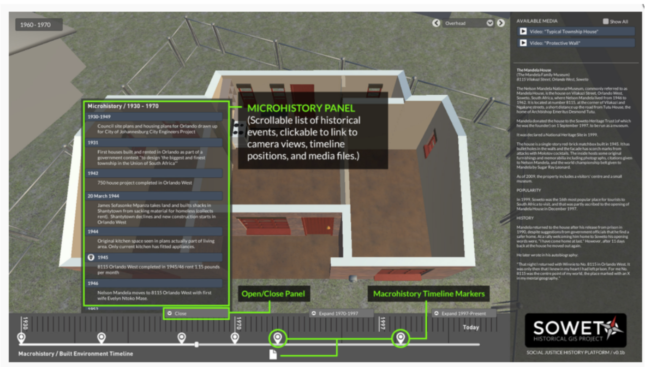


Figure 3. Emancipating Slavery Interface, University of Richmond

3D Forensics & Historical Reconstructions: Documenting Human Rights Violations During the 1976 Soweto Uprisings

Dr. Angel David Nieves

Over the past decade, scholars and community leaders have experimented with the use of new digital technologies to tell the history of the anti-apartheid movement in South Africa. Technologies now at our disposal allow us to layer victim testimony in hypertexts using multiple tools for mapping, text mining, and 3D visualizations. Digital humanities (DH) may also help analyze documentation so as to reconstruct and recover an alternative historical narrative in the face of conventional wisdom or officializing histories for the foreign tourist market. The potential layering of the many narratives also helps lay bare the messiness of archive making, the methodologies of digital ethnography, and, in particular, the endangered nature of those archives across South Africa related to the Soweto Uprisings of June 1976. As a 3D and virtual reality enabled platform (built atop the Unity engine), the Social Justice History Platform is able to represent both 2D geospatial information (such as maps, photographs, and records) and 3D representations of landscapes, locations, and 3D models of historical buildings and objects. The project combines traditional ethnographic and oral history fieldwork with 3D technologies in the pursuit of documenting past human rights violations by the former apartheid regime.


1960 - 1970
Overhead
AVAILABLE MEDIA
Show All
Video: "Typical Township House"
Video: "Protective Wall"
MICROHISTORY PANEL
(Scrollable list of historical events, clickable to link to camera views, timeline positions, and media files.)
Open/Close Panel
Macrohistory Timeline Markers
Close
Expand 1970-1997
Expand 1997-Present
1930
1970
1997
Today
Macrohistory / Built Environment Timeline
SOWETO
HISTORICAL GIS PROJECT
SOCIAL JUSTICE HISTORY PLATFORM / v0.1b

Microhistory / 1930 - 1970

1930-1949

Council site plans and housing plans for Orlando drawn up for City of Johannesburg City Engineers Project

1931

First houses built and rented in Orlando as part of a government contest "to design 'the biggest and finest township in the Union of South Africa'"

1942

750 house project completed in Orlando West

20 March 1944

James Sofasonke Mpanza takes land and builts shacks in Shantytown from sacking material for homeless (collects rent). Shantytown declines and new construction starts in Orlando West

1944

Original kitchen space seen in plans actually part of living area. Only current kitchen has fitted appliances.

1945

8115 Orlando West completed in 1945/46 rent 1.15 pounds per month

1946

Nelson Mandela moves to 8115 Orlando West with first wife Evelyn Ntoko Mase.

1952

The Mandela House
(The Mandela Family Museum)
8115 Vilakazi Street, Orlando West, Soweto

The Nelson Mandela National Museum, commonly referred to as Mandela House, is the house on Vilakazi Street, Orlando West, Soweto, South Africa, where Nelson Mandela lived from 1946 to 1962. It is located at number 8115, at the corner of Vilakazi and Ngakane streets, a short distance up the road from Tutu House, the home of Archbishop Emeritus Desmond Tutu.

Mandela donated the house to the Soweto Heritage Trust (of which he was the founder) on 1 September 1997, to be run as a museum.

It was declared a National Heritage Site in 1999.

The house is a single-story red-brick matchbox built in 1945. It has bullet holes in the walls and the facade has scorch marks from attacks with Molotov cocktails. The inside hosts some original furnishings and memorabilia including photographs, citations given to Nelson Mandela, and the world championship belt given to Mandela by Sugar Ray Leonard.

As of 2009, the property includes a visitors' centre and a small museum.

POPULARITY

In 1999, Soweto was the 16th most popular place for tourists to South Africa to visit, and that was partly ascribed to the opening of Mandela House in December 1997.

HISTORY

Mandela returned to the house after his release from prison in 1990, despite suggestions from government officials that he find a safer home. At a rally welcoming him home to Soweto his opening words were, "I have come home at last." However, after 11 days back at the house he moved out again.

He later wrote in his autobiography:

"That night I returned with Winnie to No. 8115 in Orlando West. It was only then that I knew in my heart I had left prison. For me No. 8115 was the centre point of my world, the place marked with an X in my mental geography."

Figure 4. Social Justice History Platform Interface, Soweto Historical GIS Project.

Bibliography

Bodenhamer, D., Corrigan, J. and Harris, T. (2015). *Deep Maps and Spatial Narratives*. Indianapolis: Indiana University Press.

Bodenhamer, D., Corrigan, J. and Harris, T. (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship.* Indianapolis: Indiana University Press.

Brown, V. (2016). "Narrative Interface for New Media History: Slave Revolt in Jamaica, 1760–1761." The American Historical Review, 121(1): 176-186.

Carter, B. (2010). "Virtual Harlem: Building Community." In Sara Guth and Francesca Helm (eds), *Telecollaboration 2.0: Language, Literacies and Intercultural Learning in the 21st Century.* Peter Lang, pp. 365-374.

Gallon, K. (2016) "Making a Case for the Black Digital Humanities." In Gold, M.K. and Klein, L. (eds), *Debates in the Digital Humanities.* University of Minnesota Press.

Nesbit, S., Ridge, M., and Lafreniere, D. (2013). "Creating Deep Maps and Spatial Narratives through Design." *International Journal of Humanities and Arts Computing,* 7 (October): 176-189.

Nesbit, S. (2013). "Visualizing Emancipation: Mapping the End of Slavery in the American Civil War." In Zander, J. and Mosterman, P. (eds), *Computation for Humanity: Information Technology to Advance Society.* Taylor & Francis, pp. 427-434.

Nieves, A. and Siobhan S. (2016). "Subaltern Archives, Digital Historiographies." In Singh, J. and Kim, D. (eds), *The Postcolonial World*. Routledge, pp.

The Environmental and Human Costs of DH

Alice Peaker
apeaker@brynmawr.edu
Bryn Mawr College, United States of America

Jeffrey Moro
jmoro@umd.edu
University of Maryland, United States of America

Christina Boyles
christina-boyles@uiowa.edu
University of Iowa, United States

Nicholas Weber
nicholas.m.weber@gmail.com
University of Washington, United States

Margaret Linley
mlinley@sfu.ca
Simon Fraser University, Canada

Lindsay Barbieri
lkbar@uvm.edu
University of Vermont, United States

Introduction

In her DH 2014 keynote address, Bethany Nowviskie encouraged digital humanists to "attend to the environmental and human costs of DH" (2015). These costs are sometimes accrued through acts of inaccessibility, such as through building websites that are not practical for screen readers or mobile devices. But they may also be accrued through acts of accessibility like exposing communities to unwanted surveillance through digital publications. (See, for example, the recent controversy around the digitization and open access publication of the lesbian erotic magazine *On Our Backs* raised by Tara Robertson in "Digitization: Just Because You Can, Doesn't Mean You Should", as well as Reveal Digital's response. Projects like Mukurtu are seeking to temper the open access movement by providing a platform that keeps the power of distribution and access within the hands of community members). These costs are accrued whether the intentions are deliberate or not.

In the environmental humanities, like in the digital humanities, "access" is not always a desirable goal. The actions of thousands of First Nations and Native American people, who continue to protest corporate and state access to tribal lands for the purpose of building pipelines, attest that access is both a human and environmental issue. In late October 2016, a viral Facebook campaign launched in the U.S. in response to unsubstantiated reports that the FBI was targeting Standing Rock, North Dakota protesters' locations via Facebook. Regardless of the veracity of the reports, the mass responses on social media—where individuals "checked in" *en masse* to Standing Rock—attest to a perceived privacy violation where personal data is accessed to enact environmental injustices. Access, especially human access, may well put endangered ecosystems at risk, expediting the "climate of extinction" in which, Nowviskie asserts, digital humanists work.

The papers in this session confront this "climate of extinction" both directly and obliquely.

Drawn from a diverse range of disciplinary fields and locales—including literature, media studies & archaeology, information science, and environmental science—this session interweaves examinations of *the lived ecologies of the digital* with analyses of *digital representations of lived ecologies*. Collectively, they address both the material and immaterial repercussions of digital humanities within the Anthropocene.

In the contact between humans, non-humans, and the more-than-human, this session asks us to consider the ways in which digital humanists and digital humanities projects are complicit in environmental degradation. How are digital tools leveraged to enact environmental injustices and destruction? How are they used to redress environmental injustice? How might we, in the words of The Dark Mountain Project manifesto “face this reality honestly and learn how to live with it” (“The Manifesto,” n.p)?

(Un)natural Disasters: The United States’ Racialized Response to Disaster Relief

Christina Boyles

Hurricane Matthew wreaked incalculable damage in the United States and Caribbean. According to the Weather Channel, "The eyewall may deliver the strongest, most destructive winds anyone in parts of the northeast and east-central Florida coast has seen in their lifetime. The last, and only, Category 4 hurricane to make landfall anywhere in northeast Florida or the Georgia coast was an 1898 hurricane south of St. Simons Island, Georgia" ("Hurricane Matthew a Potentially Catastrophic Category 4 or 5 Strike Ahead on Florida's East Coast; Strongest in Decades").

In fact, the severe impacts from this hurricane have led to some locations being uninhabitable for weeks or months. Governor Rick Scott encouraged residents of Florida's southeastern shore to evacuate. Other residents, like those farther north in Jacksonville or farther West near Orlando, were put on high alert. Over 1.5 million residents of these areas fled their homes ("Hurricane Matthew Strengthens as Florida Governor Urges Evacuations").

Not all residents of Florida had the means to escape the onslaught of Matthew. Many residents, particularly those living in rural and inland communities, did not have access to the resources needed to flee the coming storm. In fact, if history has anything to tell us, those without resources will suffer the most. In 1928, Florida was hit by a category 4 hurricane now referred to as the Lake Okeechobee hurricane. Although the loss of life was catastrophic—the Galveston hurricane of 1900 is the only natural disaster to have caused more American deaths—the legacy of the storm has largely been lost to history. Nicole Sterghos Brochu asserts that this is "because the vast majority of those who died were black migrant workers, segregated in life and abandoned in death" ("Florida's Forgotten Storm").

The fallout of the storm, however, has left a lasting cultural legacy in central Florida. Notably, anger has simmered for decades in West Palm Beach's African-American community over disparate memorials for black and white storm victims. Sixty-nine white victims in a segregated mass grave received personalized burial markers. In a nearby pauper's cemetery, a mass grave of 674 black victims was forgotten and left unmarked, later sharing space with a dump, a sewage plant, and a street extension ("Storm's Path Remains Scarred after 75 Years").

Government documents reveal that the racialized response to the 1928 storm was intentional. Seeking to protect Florida's burgeoning tourist industry, federal officials minimized the damages caused by the storm, even going so far as to dramatically underestimate the death toll. Since many individuals who lost their lives were transient—meaning their names and residences did not appear in census data—the government could easily downplay and even negate their existence.

To bring the stories of the storm's underrepresented victims back into our cultural memory, I created a Neatline exhibit demonstrating the loss of life the 1928 hurricane caused in both the United States and the Caribbean. To do so, I am also conducting interviews with family members of survivors and embedding their stories into the exhibit.

As Florida and the Caribbean start to recover from Hurricane Matthew, it is important to note that those living in economically disadvantaged communities will suffer the greatest from the storm’s damage. Heavily populated by black residents, these towns risk facing the same mistreatment of these residents both during the storm and in the recovery process, even recent storms—Katrina and Sandy are two prominent examples—reveal that discriminatory practices are common to disaster practices and clean-up processes.

In order to prevent a similar injustice it is crucial to point out the United States’ racialized responses to natural disasters and to focus aid efforts on the locations most impacted by the storm, many of which will be communities of color. By failing to do so, we risk contributing to a troublesome legacy of disaster relief discrimination.

No, Drones: Institutional Critique of UAS Data Accessibility and implications for Environmental Humanities

Nicholas Weber

Lindsay Barbieri

In name, pastiche isn’t a method familiar to most earth scientists. But in practice, monitoring and observing environmental phenomena is a recombinatory process not that different from pastiche of the art world; it requires collecting data via a network of remote sensing instruments, normalizing this information so that it can meaningfully interoperate, and combining different stores of data in order to reliably produce knowledge about the natural world. The process of knowledge production in the earth sciences is both techno-scientific (Haraway, 1997), in the sense that sources of data are historically and locally situated, but also highly contingent on evidential cultures (Collins, 1998;

Baker, 2011) in that collecting, finding and using data is mediated by the background practices of an academic department, scientific discipline, or research program.

Compare our generic description of knowledge production in environmental monitoring to Louise Lawler's recent series titled 'No Drones' – a critique of contemporary art institutions. Lawler created the series through a pastiche process of photographing private art collections, commissioning drawings of her photographs, and then digitizing the illustrations as vector-based images that are magnified and printed on adhesive vinyl (see Image 1 below). For an unassuming viewer, the result is a powerful reflexive image that is both full of layers of symbols and yet stripped down to a single black-and-white outline of complex ideas. For an audience that can unpack the provenance of Lawler's images this work also creates a self-reflexive form of institutional critique of how highly priced and commoditized artworks serve as cultural markers of wealth, privilege, and accessibility (Nixon, 2014).

This paper attempts to use the same form of institutional critique as Lawler (Fraser, 2005) in discussing the emergence of Unmanned Aerial Systems (UAS) in environmental monitoring. Our goal is to demonstrate that with reflexivity the process of producing knowledge about environmental phenomena with UAS data can shed considerable light on markers of wealth, privilege and accessibility at play in contemporary informatics-driven science. Further, we argue that by creating 'legible' provenance information, UAS data can have a dramatic impact on the burgeoning practices of environmental humanities (Castree, 2014).

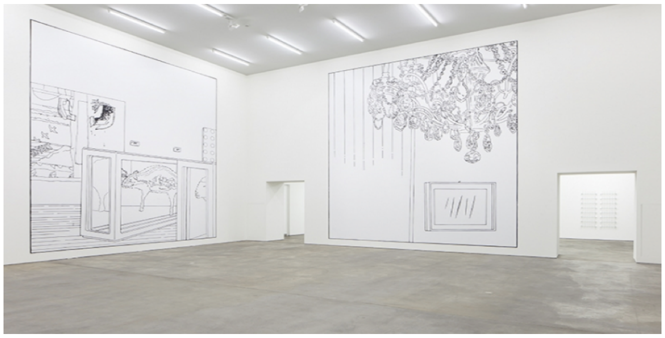


Figure 1: Louise Lawler's installation 'No Drones' at Sprüth Magers, Berlin 2015. Pictured left is 'Dots and Traces' - A drawing by Jon Buller that has been digitized and printed on a vinyl adhesive. Buller's sketch is a direct copy of photo taken by Lawler, which is itself a sculpture by Damien Hirst. (Image via Sprüth Magers)

The use of unmanned aerial systems (UAS) as tools for collecting innovative remote sensing data and geospatial imaging is increasing. Recent technological advances present communities with an opportunity to use UAS to push boundaries in data collection. The recent drop in costs combined with the advancement in lightweight technology mean UAS could become a ubiquitous means of collecting scientific data (Dunbabin and Marques, 2012). UAS are predominantly used for imagery, however, given their advantages over traditional data capture methods (higher temporal and spatial resolution, new area access, etc.) novel sensors are being explored and UAS use has opened doors for both researchers and communities interested in monitoring environment, landscapes, community vulnerability and disaster response.

We present on three important topics in UAS-based collection of environmental data (1) the results of a survey of UAS use in the earth sciences (flight platforms, data and metadata standards, and data sharing practices) and how this may shape the ability of communities to access these technologies (2) a discussion of the experience of unwanted surveillance and how drones as tools could shift the equity of surveillance (self surveillance / community surveillance) and power dynamics. (3) a discussion of how drone data collection, with unprecedented ability to capture data on environment and landscape, may shift responses of organic beings to their physiographic surroundings – a topic of growing importance for the field of environmental humanities.

The English Lake District and 'World Ecology'

Margaret Linley

In his *Guide to the Lakes* (1835), William Wordsworth famously condemns changes occurring in the Lake District since the late eighteenth century, changes accelerating ironically because visitors were flocking there from all parts of England to see the landscapes Wordsworth celebrated in his nature poetry. Wordsworth is induced to speak out "at length" by a "wish to preserve the native beauty" of the district against the transformations that seem so rapidly to be taking place. The new mobility and accessibility enabled by improved transportation and communication, however, are only part of the problem. The "invention and universal application of machinery" are, for Wordsworth, equally to blame. Rather than simply argue for isolation, Wordsworth's solution to this conundrum of access and preservation is to use the mass print genre of the travel guide as an educational platform for cultivating environmental consciousness toward a reconstitution of the Lakeland as public space: "a sort of national property, in which every man has a right and interest who has an eye to perceive and a heart to enjoy" (*Guide*).

Early in the *Guide* Wordsworth describes the process by which "nature is indebted to the hand of man," its state and appearance the outcome of historical and social processes. In so doing, he complicates a strain of environmental writing, interestingly often attributed to the Romantic legacy, which differentiates human activity from "the natural," a habit of thought that, according to ecocritics such as Jason W. Moore (*Capitalism in the Web of Life*), ultimately mystifies the role of capitalism in resource exploitation and climate change. Yet perhaps even more provocative is the way Wordsworth analyses the history of travel in the Lake District as essentially a form of colonization and insists, additionally, that the operations of such modernized forces of power on nature demand that readers think ecologically.

This paper will explore how the problem of access Wordsworth identified dovetails with concerns raised by scholars about the environmental and human costs of digital humanities (such as Nowviskie, Parikka, and Haraway, among many others) through the specific example of *Lake District Online*, a digital research project based on Simon Fraser University's collection of 260 illustrated rare books about the English Lake District, including many maps and historical specimens of ornate book bindings, illustrations, and photography, spanning 300 years (1709-2000) with a concentration in the nineteenth century. As Elizabeth DeLoughrey and George B. Handley argue, historicization has been a primary tool of postcolonial studies and it is central to our understanding of land and, by extension, the earth. In order to continue to engage a historical model of ecology and an epistemology of space in time, we must carry these concepts, not least of all the spatial imaginary made possible by the experience of place, into the dialogue around access to emerging digital environments.

Guided by this premise, *Lake District Online* has produced several freely accessible research, educational, and public engagement resources: 1) an extensible bibliographic research database for linking with other open databases (such as Wikipedia), comprising a metadata framework for searching and indexing bibliographical, biographical, and critical information about the books as well as their contents; 2) an experimental prototype of a teaching and learning platform, *Reading Up Close and At a Distance*, designed to enable students to interpret literature interactively on different scales; 3) a co-curated public exhibition with the Wordsworth Trust, *Wordsworth Country: From the English Lake District to the Pacific Northwest*, launched simultaneously at the Wordsworth Museum (UK), SFU Library and Special Collections (BC), and online; and 4) a corpus of high-resolution digital images and text files based on the SFU Lake District rare book collection. To approach these resources through the paradigm of colonial violence that Wordsworth identifies at work in the regional landscapes of Cumbria is also to engage empire building and environmental histories as highly mobile, flexible, and mutually constitutive. Moreover, as Wordsworth implies, access - and the related concept of openness - is a question of spatiality and especially of boundaries, exclusions, and limitations entailed in the politics and ethics of place making. Focusing especially on the significance of the present location of the Lake District collection in Vancouver, British Columbia, this paper will reflect on colonial legacies of organizing, producing, and accessing nature in the context of the digital humanities.

Decaying Plastic Play: *Flappy Bird*'s Hacked Afterlife as Media Archaeological Praxis

Jeffrey Moro

On March 28th, 2016, prolific YouTube streamer SethBling posted a video demonstrating how, using only timed button presses and graphical glitches present in the console original, he injected three hundred and thirty-one new bytes into the seminal 1990 Super Nintendo Entertainment System (SNES) platforming game Super Mario World—bytes corresponding to the source code of the 2013 viral iPhone game phenomenon Flappy Bird. The hack allows users to play a fully functional port of Flappy Bird within Super Mario World, grafting the former's computational logic into the latter's graphics. The choice of games here is striking; while Super Mario World has been re-released across a variety of hardware platforms—to say nothing of Mario's cultural ubiquity—Flappy Bird remains a touchstone for its inaccessibility, both in its frustratingly difficult design and the fact that in February 2014, its creator pulled it from all platforms, citing concerns that its addictiveness ruined people's lives (Nguyen). SethBling's hack then produces a chimeric object; a hybrid of plastic, logic, and time; the ghost of one game haunting the shell of another. This haunting constitutes, as this short paper argues, a regenerative practice: a "circuit bending," to draw on a theoretical and practical challenge from Garnet Hertz and Jussi Parikka, that engages the "archive" of dead (whether by accident or design) computational media not only within artistic traditions of remix, tinkering, and collage, but also with a media archaeological eye towards circulations of technological waste, supply chains, and resource extractions ("Zombie Media").

In characterizing this haunting as "regenerative," I engage interdisciplinary work in the environmental humanities and media archaeology alongside tactics from games and software studies. In a post-400ppm world, questions of digital media's emergence from and contribution to the feedback loops of climate crisis within industrial capitalism loom large. In the digital humanities, Bethany Nowviskie's oft-cited keynote at DH2014 in Lausanne provides a cri de coeur, alongside emerging conversations from the GO::DH working group on minimal computing from the same conference, for thinking through the ecological circulations and impacts of DH technologies and practices—what Nowviskie, drawing from Steven Jackson, calls the challenge of "broken world thinking" (5). Media studies as a field has also begun to take up broader questions of climate, geology, infrastructure, and the Anthropocene in theoretical and material studies of digital technologies, whether under the rubric of the "nonhuman turn" (ex. Grusin, Richard, ed. The Nonhuman Turn, U of Minnesota P, 2015); geological and elemental media histories (ex. Parikka, Jussi. A Geology of Media. U of Minnesota P, 2015; Peters, John Durham. The Marvelous Clouds: Toward a Philosophy of Elemental Media. U of Chicago P, 2015); computational infrastructure (ex. Bratton, Benjamin. The Stack: On Software and Sovereignty. MIT P, 2015; Starosielski, Nicole. The Undersea Network. Duke UP, 2015; Blum, Andrew. Tubes: A Journey to the Center of the Internet. Ecco, 2013); or sustainability and waste (ex. Gabrys, Jennifer. Digital Rubbish: A Natural History of Electronics. U of Michigan P, 2011; Acland, Charles, ed. Residual Media. U of Minnesota P, 2007; Scanlan, John. On Garbage. Reaktion Books, 2005). These diverse approaches, only a small fraction of such work, share a common reading of computational technologies both relying on and contributing to a material reorientation of human/Earth relations, whether through open-pit mining, ozone depletion, or undersea cable networking—to say nothing of new social relations made possible by the connective potentiality of such technologies.

It is no coincidence that SethBling's hack emerges at the same time as academic and environmentalist communities' explorations of technological obsolescence, material production, and sustainability/waste grow more prominent. Nor is he working in isolation: his hack is only one within of a subculture of videos exploring different ways to hack, deform, and manipulate "classic" (almost always a euphemism for "obsolete") video games. Part of these players' engagement with obsolete video games and platforms is practical: the relative (to contemporary platforms) simplicity of the hardware and coding allows for more granular engagement with the technologies themselves. But the choice of Flappy Bird and Super Mario World reveals multiple valences of computational obsolescence and decay: a game forcibly disappeared emerges within one that, through the cooperative/coercive machinations of nostalgia and capitalism, endures. Moreover, the hack itself is a kind of deliberate decay, one that deforms the digital object until it becomes pliable, manipulable, and inscribable. Much like the acrylonitrile butadiene styrene (ABS) that comprises its cartridge shell, the game becomes plastic: transformable, recyclable, and pliable under the right (industrial) conditions.

At its root, this hack plays (in multiple senses of the word) with the roots, tendrils, and growths of computational memory. It reveals "memory" as a function of physical hardware, the encoded rhetoric of software, and the transmission of shared culture. Wendy Hui Kyong Chun, in her 2008 book Programmed Visions, offers the idea of software's "enduring ephemerality," computational media's capacity to "remember" through material regeneration—continual acts of writing and rewriting across electrical charge and silicon (148). Through this frame, SethBling's hack is rewriting, and its ingenious incarnating and recycling of Flappy Bird's three hundred and thirty-one bytes embodies Chun's claim that "what is not constantly upgraded or 'migrated' or both becomes unreadable. . . . The experiences of using—the exact paths of execution—are ephemeral. Information is 'undead': neither alive or dead, neither quite present or absent." (148). Hertz and Parikka deploy "zombie" as a metaphor for reinvigorated dead media to similar ends, observing that the materials constituting software—the plastic, the rare earth metals—are also subject to regimes of ephemerality (150–53). ABS, in its petroleum-based non-biodegradability, may be effectively immortal, but the systems through which humans articulate its use are subject to entropic decay. Only the ephemerality endures.

This paper closes by offering SethBling's hack and the broader gaming subcultures of glitch play to which it belongs as artistic responses to computational obsolescence and decay. Chun's "enduring ephemerality" becomes artistic praxis much as Hertz and Parikka offer circuit bending as a media archaeological arts method. It extends the material concerns of media studies' Anthropocenic turn to code and its cultures, and offers these deforming/reforming modes of play as potential sites of resistance for an ecologically-engaged digital humanities practice.

Bibliography

Baker, J. D. (2011). Tradition and toxicity: evidential cultures in the kava safety debate. Social studies of science, 0306312710395341

Castree, N. (2014). The Anthropocene and the environmental humanities: extending the conversation. Environmental Humanities, 5(1), 233-260

Chun, W. H. K. (2011). *Programmed Visions: Software and Memory*. MIT P

Collins, H. M. (1998). The meaning of data: Open and closed evidential cultures in the search for gravitational waves 1. American Journal of Sociology, 104(2), 293-338.

Dunbabin, M., & Marques, L. (2012). Robots for environmental monitoring: Significant advancements and applications. IEEE Robotics & Automation Magazine, 19(1), 24-39.

Fraser, A. (2005). From the Critique of Institutions to an Institution of Critique. Artforum, 44(1), 278.

Haraway, D. J. (1997). Modest– Witness@ Second– Millennium. FemaleMan– Meets– OncoMouse: Feminism and Technoscience. Psychology Press Nixon, M. (2014). Louise Lawler: No Drones. October, 20–37. https://doi.org/10.1162/OCTO_a_00164

Hertz, G.,and Parikka, J. (2015) "Zombie Media: Circuit Bending Media Archaeology into an Art Method." Appendix. *A Geology of Media*, by Jussi Parikka. U of Minnesota P, 2015, pp. 141–53.

Minimal Computing. (n.d.) "Minimal Computing: a working group of GO::DH." Accessed Oct 27, 2016, http://go-dh.github.io/mincomp/

Nguyen, L. A. (2014) "Exclusive: Flappy Bird Creator Dong Nguyen Says App 'Gone Forever' Because It Was 'An Addictive Product.'" Feb 11, 2014. *Forbes*, accessed Oct 27, 2016, http://www.forbes.com/sites/lananhnguyen/2014/02/11/exclusive-flappy-bird-creator-dong-nguyen-says-app-gone-forever-because-it-was-an-addictive-product/

Nowviskie, B. (2015). "Digital Humanities in the Anthropocene." *Digital Scholarship in the Humanities*, special printing, pp. 1–12, doi:10.1093/llc/fqv015.

SethBling. (2016) "SNES Code Injection -- Flappy Bird in SMW." Mar 28, 2016. *YouTube*, accessed Oct 27, 2016, https://www.youtube.com/watch?v=hB6eY73sLV0

Humanidades Digitales en Iberoamérica: desafíos institucionales para su desarrollo y consolidación

Adriana Álvarez Sánchez
adralvsan@gmail.com
Universidad Autónoma Nacional de México, Mexico

Miriam Peña Pimentel
miriampeñapimentel@gmail.com
Universidad Autónoma Nacional de México, Mexico

Esteban Romero Frías
eromerofrias@gmail.com
Universidad de Granada, Spain

Virginia Brussa
virbrussa@gmail.com
Universidad Nacional de Rosario, Argentina

Paola RicaurteQuijano
pricaurt@itesm.mx
Tecnológico de Monterrey, Mexico

Enedina Ortega
enedina.ortega@gmail.com
Tecnológico de Monterrey, Mexico

Cristóbal Suárez-Guerrero
Cristobal.Suarez@uv.es
Universitat de València, Spain

Ana Teresa Morales Rodríguez
ateremora@gmail.com
Universidad Veracruzana, México

Alberto Ramírez Martinell
ateremora@gmail.com
Universidad Veracruzana, México

Resumen

En las universidades iberoamericanas las Humanidades Digitales están construyendo un campo emergente con diversas iniciativas a nivel institucional. Las instituciones de educación superior poseen características diversas que las distinguen por sus recursos humanos y financieros, sus infraestructuras físicas, tecnológicas y administrativas, sus modelos educativos y sus programas de estudio. Esta heterogeneidad genera condiciones diversas para la consolidación del campo. Sin embargo, aunque los contextos varíen, nos encontramos con elementos comunes en esas estructuras institucionales que obstaculizan el desarrollo de las Humanidades Digitales y, a la vez, pliegues o espacios liminales que posibilitan su avance. Este panel tiene el propósito de a) presentar un marco de referencia general sobre las HD en Iberoamérica, b) realizar un análisis de los desafíos como campo y un balance de los factores determinantes para el avance de las HD en español, c) identificar las estrategias que han sido aplicadas para la expansión de las HD en la región y, d) proponer acciones en red y mecanismos dentro de las universidades que permitan la consolidación del campo, a través de la práctica docente, la investigación, el desarrollo de capacidades, la formación de recursos humanos y la vinculación social de las Humanidades Digitales en el marco de los principios de una cultura digital crítica y libre.

Descripción

En Iberoamérica existen instituciones de educación superior con características diversas que las distinguen por sus recursos humanos y financieros, sus infraestructuras físicas, tecnológicas y administrativas, sus modelos educativos y sus programas de estudio. Esta multiplicidad de contextos genera dinámicas y espacios diferenciados para la consolidación de las Humanidades Digitales en la región. Dentro de este contexto institucional variado, las HD en español están construyendo un campo emergente que posee características que lo distinguen de las experiencias en otras regiones y otras lenguas. Sin embargo, al abrir un línea de trabajo nos encontramos con múltiples elementos de esas estructuras institucionales que obstaculizan su avance. La formación de profesionales y el desarrollo de proyectos digitales se enfrentan a una serie de problemáticas dentro de las que se encuentran decisiones administrativas, el desconocimiento sobre el campo, las carencias en infraestructura o de mecanismos para que las universidades incluyan nuevas temáticas en el currículo a nivel de licenciatura, posgrado o como líneas de investigación.

Sostenemos que las instituciones cuentan con una serie de elementos estructurales que, a su vez, forman parte de un sistema de educación pero también de producción de conocimiento, por ello enumeramos aquí algunos de los elementos externos e internos del contexto en el que buscamos consolidar las HD en Iberoamérica con el fin de presentar un diagnóstico y una propuesta que permita a las instituciones tomar decisiones estratégicas si buscan transitar a un nuevo paradigma.

- Administración: las instituciones cuentan con un aparato administrativo que, en ocasiones, representa un obstáculo para el desarrollo de nuevas propuestas. El planteamiento de nuevos contenidos y conocimientos debe sujetarse a la estructura administrativa, lo cual puede permitir o impedir el desarrollo de las HD.
- Evaluación de la producción académica: aunque no es un problema nuevo, la presencia de medios

digitales complejiza la valorización de los productos intelectuales, sólo tangibles en productos ya validados, como los libros o los artículos, sin embargo, hoy es posible desarrollar productos en soportes y formatos diferentes para los cuales no existe aún un mecanismo de validación o el que existe no ha sido adaptado para estas nuevas formas de producir conocimiento.

- Modelo educativo, planes de estudio y propuesta pedagógica: los distintos modelos educativos, planes de estudio o propuestas pedagógicas, sea cual sea el programa o nivel académico, obedecen a la división tajante del conocimiento en disciplinas, ello impide en la mayoría de los casos, impartir asignaturas que integren nuevas visiones pedagógicas, formas de trabajo, objetos de estudio, teorías y métodos de investigación o incluso romper con la compartimentación disciplinaria y la ampliación de entornos de aprendizaje
- Comunidad y cultura digital: la formación de profesionales depende no sólo del interés de los docentes, sino también de los propios estudiantes. Ambos grupos reconocen la necesidad de adquirir nuevos conocimientos y competencias digitales pero en muchas ocasiones únicamente desde un enfoque instrumental, lo cual no implica una reflexión sobre el impacto de medios digitales y de internet en el desarrollo de sus disciplinas.
- Infraestructura tecnológica: el acceso en términos de conectividad, hardware y software.
- Transdisciplinar: resultado de la fragmentación y especialización disciplinar dentro de las estructuras universitarias, el trabajo colaborativo interdisciplinar y transdisciplinar es aceptado con cierta reticencia, dado que su desarrollo debe apegarse a las estructuras académicas y administrativas ya establecidas.
- Financiamiento: la investigación requiere de recursos económicos para la realización de proyectos pero también para la formación de los humanistas. Los aspectos mencionados se encuentran dentro de un contexto más amplio que reafirma la división estricta entre disciplinas, lo cual incide en el financiamiento para proyectos digitales desde la perspectiva de las HD. El acceso a recursos internacionales es limitado puesto que requiere de redes de colaboración y capacidades con las que no necesariamente cuentan los investigadores. Políticas públicas.

A pesar de que esta estructura institucional no siempre es favorable para el avance de las HD, se han desarrollado iniciativas en las universidades que van desde la organización de eventos de HD, la publicación académica de revistas especializadas, la creación de plataformas, la vinculación con otras iniciativas en el marco del movimiento de acceso abierto (MOOC, bases de datos, bibliotecas digitales, Wikipedia); hasta la implementación de laboratorios como parte de la estrategia para abrir o consolidar la práctica de las HD. El panel sobre Humanidades Digitales en Iberoamérica esbozará un estado de la cuestión y, a través de las diversas experiencias institucionales de los panelistas, buscará identificar los principales desafíos, prácticas exitosas y propuestas para generar mecanismos dentro de las universidades que permitan avanzar en el desarrollo del campo, la práctica docente, la investigación, el desarrollo de capacidades, la formación de recursos humanos y la vinculación social de las Humanidades en el marco de una cultura digital crítica y libre.

Introducción al panel

El panel muestra diversos intereses, aunque con un objetivo común: conocer las estrategias que han sido aplicadas para la expansión de las HD en Iberoamérica. Aunque existen algunos mapas y registro de proyectos, nuestro interés es entrar a profundidad en el análisis de algunos factores que consideramos determinantes para el avance de las HD en español. Como punto de partida, el panel presenta un marco de referencia general sobre las HD en Iberoamérica como introducción a los casos y cierra con un balance de la situación y los desafíos como campo con el propósito de identificar continuidades y fortalecernos a través de estrategias compartidas.

El panel aborda estudios de casos relevantes que reflejan la labor de las instituciones Iberoamericanas en las se han desarrollado proyectos e iniciativas para extender las HD, que han generado ya redes de colaboración y un sentido de pertenencia y reconocimiento mutuo. Este panel como resultado de esas redes de cooperación interinstitucional e internacional busca consolidar una apuesta conjunta, generar un espacio de discusión abierto y dar visibilidad al trabajo heterogéneo que se realiza en Iberoamérica. Si bien no se mapean todos los casos de HD en Iberoamérica, hay una voluntad explícita de incorporar, reconocer y rescatar las iniciativas de la región. Esperamos que a partir de la oportunidad de una discusión pública, podamos, en un futuro próximo, integrar académicos y experiencias de otras instituciones y países.

Desde Iberoamérica la comprensión de las HD es más amplia que en el mundo anglosajón, lo que muestra también un desafío en términos académicos, administrativos y de práctica de las HD. Consideramos relevante destacar nuestros contextos específicos de acción para luego identificar las estrategias más efectivas para avanzar en el campo, no únicamente a través de proyectos, sino a través de la formación de estudiantes, el desarrollo de competencias entre los investigadores y el incentivo de formación a futuro de una infraestructura colaborativa. Sostenemos que es necesario conocer el estado de la cuestión desde los aspectos más macro hasta los más micro.

La inclusión de temas sobre tecnología, educación, cultura digital y transformación institucional en el panel responde a la necesidad de presentar la pluralidad de estrategias, proyectos e iniciativas que buscan ampliar las posibilidades de institucionalización de las HD, generando condiciones básicas para avanzar en la reflexión, el planteamiento de nuevos proyectos y la consolidación del campo en la región.

Las HD en los programas de Licenciatura de Humanidades. Estado de la cuestión

Adriana Álvarez Sánchez

Miriam Peña Pimentel

La división del conocimiento en Áreas ha permitido un alto grado de especialización disciplinaria, sin embargo, la inclusión de nuevos objetos de estudio - incluidos los digitales - requiere aproximarse y, en ocasiones, transitar hacia otras áreas del conocimiento, para ello es necesario que los estudiantes en formación conozcan no sólo los principios teóricos y técnicas de otras disciplinas, sino que lleven a la práctica esa interdisciplinariedad. Ello podría conseguirse si durante la licenciatura (bachelor) cuentan con, al menos, una experiencia de trabajo colaborativo con estudiantes de otras áreas. La inter y transdisciplina son consideradas características inherentes de las HD; por ello, la réplica de esta forma de trabajo podría ser incorporada a los planes de estudio actuales. El uso de recursos y herramientas digitales facilitan la colaboración entre disciplinas; sin embargo los planes de estudio actuales no consideran esta vinculación como un requisito, por el contrario, la tendencia en el "entrenamiento" de los estudiantes, tiende hacia el individualismo y la hiper-especialización; la única posibilidad de conexión con otras áreas del conocimiento (humanístico, social o científico) es la posibilidad de cursar materias optativas en otras facultades o de otras disciplinas -siempre y cuando se demuestre cumplir con los pre-requisitos de cada materia, condición que se cumple con dificultad entre más "alejadas" estén las disciplinas-; la gestión o desarrollo de proyectos en pequeña o gran escala, no se enfocan en entablar procesos de investigación colaborativos, las estructuras jerárquicas se mantienen y la cadena de aprendizaje mantiene el status quo; perpetuando el modelo "tradicional" de trabajo. Es importante señalar que esta propuesta no ataca o desprestigia la estructura ni las formas tradicionales de la academia, por el contrario, busca ofrecer una serie de recomendaciones que en convivencia enriquezcan la experiencia universitaria en los diferentes niveles de acción: estudiantado, profesorado, investigación y administración.

En esta ponencia nos centraremos en dos disciplinas concretas: Letras Hispánicas e Historia.

La Licenciatura en Filología Hispánica tiene diferentes connotaciones dependiendo de la Universidad que imparta el programa de estudios (Lengua y Literatura Hispánicas, Letras Hispánicas, Literatura Latinoamericana, etc.); el nombre de la carrera denota la tendencia a la que se enfoca cada licenciatura, desde el formato más general (lingüística y literatura), hasta el de enfoques delimitados (literatura latinoamericana); sin embargo, para fines prácticos, los egresados de cualquier modalidad de esta licenciatura, cuentan con una instrucción similar a lo largo de 8 semestres. A pesar de las diferencias en los planes de estudio que aquí se presentarán, en su mayoría coinciden con una falta evidente de apertura hacia la inclusión de propuestas "alternativas"; en el caso de la UNAM, por ejemplo, el plan de estudios de la Licenciatura en Lengua y Literatura Hispánicas es de 1999; razón por la cual se entiende innecesario incluir una línea pedagógica hacia las Humanidades Digitales (o hacia la literatura electrónica o las "nuevas tecnologías"); sin embargo, en planes más contemporáneos (2012 en el caso de la Universidad Iberoamericana), encontramos faltas similares que, a juicio de las autoras de esta propuesta, de incorporarse a las prácticas pedagógicas podrían dar parte a una serie de programas interdisciplinares con tendencia a las Humanidades Digitales y la Pedagogía Digital; tendencias que consideramos necesarias para la innovación académica en las regiones periféricas.

La Licenciatura en Historia de la Facultad de Filosofía y Letras, de la Universidad Nacional Autónoma de México (UNAM) es un programa centrado en la historiografía como eje rector de la formación de un historiador al que se busca formar para la investigación, la docencia y la difusión. No obstante, que el programa fue creado a finales de los noventa, en su contenido únicamente se menciona en una ocasión Internet como parte de los métodos de búsqueda bibliográfica (Plan de Estudios de la Licenciatura en Historia, 1999, Vol. I, p.67).

Para realizar el diagnóstico planteado en el panel, se llevará a cabo una revisión crítica de los planes de estudio de estas Licenciaturas que se ofertan en la ciudad de México: Puesto que ambas licenciaturas tienen presencia en la mayoría de las universidades capitalinas, se decidió revisar los planes de estudios de las mismas instituciones: la Facultad de Filosofía y Letras (Escolarizado), UNAM; Universidad Autónoma Metropolitana, Iztapalapa, y Universidad Iberoamericana, esta última de carácter privado. El objetivo, además de comparar los programas, es identificar "áreas de oportunidad" para plantear una propuesta para impulsar la interdisciplinariedad y el trabajo colaborativo a través de medios digitales, lo cual implicará considerar temas como el acceso, la infraestructura, etc., es decir, el contexto institucional de cada caso. Se considerarán las experiencias institucionales que la autora de la ponencia, en colaboración con la Dra. Miriam Peña Pimentel, han tenido como profesoras del Seminario Taller Especializado Humanidades Digitales (2014-2015) e Historia, e Historia Digital (2016) dentro de la Licenciatura en Historia, UNAM; el Seminario de Humanidades Digitales (2013-2016) y la Optativa: Humanidades Digitales en los estudios literarios (2016-2 y 2017-1); además de esfuerzos semejantes llevados a cabo en otras instituciones mexicanas, sin dejar de lado el contexto mundial. Todo ello con la intención de presentar un diagnóstico conjunto que permita crear un modelo para introducir las HD a nivel licenciatura en México.

La evaluación del conocimiento en la sociedad digital. Progresos para una alternativa en Humanidades Digitales

Esteban Romero Frías

Weller (2011), autor del libro The Digital Scholar, emplea el término Digital Scholarship para referirse al conjunto de actividades académicas facilitadas por las nuevas tecnologías. Un conjunto de cambios profundos que, para este autor, son producto de la convergencia de tres elementos: lo digital, lo conectado en red y lo abierto. Las transformaciones digitales producidas en el espacio de las Humanidades, denominadas Humanidades Digitales, representan un caso de especial interés para observar el desarrollo de estas prácticas, en muchas ocasiones conflictivas. Goodfellow, en un artículo de 2013 titulado "Scholarly, digital, open: an impossible triangle?", señalaba justamente que lo abierto, lo académico y lo digital, constituyen categorías difícilmente compatibles tal y como se consideran en la actualidad. Ciertamente, en el ámbito de las prácticas digitales vinculadas al conocimiento, el conjunto de categorías que hay que tener en cuenta es muy diverso, existiendo un continuo en una pluralidad de dimensiones en las que podemos situar los distintos casos con los que nos enfrentamos. Algunas de estas categorías son, entre otras: lo académico/no académico, lo digital/analógico, lo abierto-cerrado. Junto a esto surgen en el ámbito de la sociedad digital nuevas formas de generación y difusión del conocimiento que no son reconocidas por los sistemas de evaluación tradicionales centrados en el impacto bibliométrico. Más allá aún, la apertura que generan los medios sociales hace preciso medir otros impactos de la vida académica aparte del vinculado a la investigación. Esta problemática se agrava particularmente cuando nos referimos a las Humanidades Digitales, herederas de modelos tradicionales asentados durante muchos años. Para un efectivo desarrollo y consolidación de las Humanidades Digitales es preciso identificar y evaluar los nuevos artefactos digitales académicos que en muchos casos constituyen el principal resultado de un proyecto o de la carrera de un académico.

El proyecto Knowmetrics - evaluación del conocimiento en la sociedad digital, financiado en julio de 2016 por la Fundación BBVA en la línea de Humanidades Digitales, desenvuelve precisamente su trabajo en dos líneas. En la primera, con una perspectiva micro, se centra en la identificación de investigadores en Ciencias Sociales y Humanidades Digitales, en la elaboración de una taxonomía de artefactos digitales académicos y de indicadores para evaluarlos y en la propuesta de informe integrado de los impactos de la vida académica para humanistas digitales, incluyendo diversas dimensiones, desde la investigación a la implicación social. En la segunda línea, con una visión macro, trabaja en la evaluación del impacto digital del conocimiento en la universidades a través de diversas aproximaciones: a través de un índice Knowmetrics que se apoye en el trabajo anterior, a través de altmetrics y de medios propios de Internet, como es el caso de Twitter.

La aportación en la mesa redonda que se propone permitirá dar a conocer los avances efectuados por dicho proyecto cuando se encuentre en el ecuador de su ejecución en 2017.

Repensar la producción de conocimiento, las instituciones y las humanidades digitales

Virginia Brussa

Paola Ricaurte

Enedina Ortega

En nuestro actual contexto, son frecuentes los cuestionamientos con respecto a la labor de las universidades, el sentido de la producción académica y su vinculación con los complejos problemas que aquejan a la sociedad. De acuerdo con Heleta (2016) los académicos no se encuentran perfilando los debates públicos. Menciona que anualmente se publica un millón y medio de artículos en revistas académicas, que son en su mayoría ignorados por la comunidad científica. En el caso específico de las humanidades, menciona que el 82% de los artículos no se cita ni una vez. Ello enfatiza una problemática que va más allá del acceso, también denota la necesidad de acción y agenda amplia de lo "abierto".

En una entrevista realizada a Saskia Sassen (Torres, 2016) la socióloga destaca que el mundo académico no está respondiendo a las particularidades del momento. Los académicos, menciona Sassen, se "instalan" en zonas de confort: su carrera académica, sus publicaciones, el uso de categorías dominantes, y no arriesgan, ni en la comprensión profunda y problematización de los conceptos para abordar fenómenos contemporáneos, ni en abrir nuevas fronteras de investigación. Kathleen Fitzpatrick (2011) utiliza el concepto de obsolescencia como una categoría pertinente para dar cuenta de una serie de condiciones culturales asociadas con el sistema actual de producción y difusión del conocimiento en horizonte tecnocultural: la revisión por pares, las nociones de autoría, la categoría del texto y el papel de la universidad.

Esta situación pone de relieve al menos tres problemáticas que nos interesan: por una parte, la necesidad de reconfigurar el papel (y la estructura administrativa) de las instituciones educativas para que den respuestas a los problemas sociales desde nuevas aproximaciones y espacios de "encuentro" ; por otra, transformar los sistemas de producción de conocimiento; y por último, reflexionar sobre el lugar de las humanidades (específicamente digitales) en este escenario. En este último punto, enfatizar el estado de situación de modalidades de

Labs y la creación de herramientas digitales en el seno de nuestros espacios de HD.

Esta propuesta tiene como propósito ofrecer un panorama sobre a) un modelo de transformación institucional; b) una estrategia que permite reconceptualizar la producción de conocimiento desde la academia; y c) una apuesta para promover visiones alternativas sobre el trabajo de los académicos en las humanidades (digitales).

Si buscamos el desarrollo y la consolidación de las HD como campo en Iberoamérica, es necesario fortalecer espacios que fomenten modalidades democratizadoras y colaborativas en la producción de conocimiento, frente a la institucionalidad, estructura y formas pedagógicas propuestas en el seno del ámbito académico actual. Sostenemos la caducidad y el agotamiento de las instituciones sociales (universidad, partidos políticos, dependencias gubernamentales) concebidas como sistemas cerrados y autocontenidos, desconectados de la ciudadanía. Apelamos a la conformación de extituciones: formas organizacionales abiertas a la participación ciudadana como un aspecto fundacional de su constitución y operación.

La universidad debe reconfigurarse bajo este modelo para tener cabida en las nuevas dinámicas sociales. En este sentido, planteamos que las Humanidades Digitales no deberían estar exentas de estos debates y desafíos si se encuentran inscritas en los valores de colaboración, apertura e interdisciplinariedad. Manifestamos la necesidad de una opción crítica y pública en las humanidades que se hacen desde el sur global. Por esta razón, consideramos pertinente repensar los modelos triple hélice que consolidan los lazos entre universidad, gobierno y sociedad civil.

Frente a ello, nuestra propuesta se basa en presentar un planteamiento sobre la vocación y orientación de la humanidades en nuestra región, que contempla una serie de propuestas relativas a nuevas visiones, temáticas, procesos y proyectos. En relación con el impulso de los debates sobre la universidad de hoy, es menester hacer mención a la creación de nuevos espacios, metodologías y modalidades de producción del conocimiento, a la conformación de otras redes con comunidades y organizaciones, a la incorporación de esquemas flexibles y alternativos de colaboración, a las formas de mediación y las conversaciones en distintos niveles de lo “abierto”: institucional (gobiernos, universidades, ONG), temáticos (gobiernos abiertos, datos abiertos, ciencia abierta, ciencia ciudadana), procesos y herramientas. Consideramos que desde las universidades es importante rescatar la vocación de experimentación y la generación de conocimiento en abierto para transformar la cultura académica, el entorno y las comunidades. Las humanidades tienen una oportunidad invaluable de incidencia en este proceso.

Proponemos por tanto, líneas de acción que podrán aportar a dichos debates y problemáticas que enfrenta la universidad, así como el impacto que ello conlleva para la materialización de las Humanidades Digitales críticas y públicas en el ámbito regional.

La reconfiguración social de aprender en red

Cristóbal Suárez-Guerrero

Paola Ricaurte Quijano

La educación de hoy se piensa y hace en internet. No obstante, el perfil y el papel de los agentes educativos en los entornos de aprendizaje en red es algo que cuesta definir no solo por la multiplicidad de agentes que hay, sino porque el paradigma dominante de “experiencia de aprendizaje” está sujeta a la relación unidireccional de profesor-estudiantes. Esta matriz de relación educativa entra en crisis cuando hablamos de interacción educativa en internet. Por tanto, más allá de las herramientas tecnológicas es preciso preguntarse ¿qué implica hablar de lo social cuando nos encontramos aprendiendo en internet? Para dar respuesta y entender la importancia de lo social en el aprendizaje hay que apelar a una pregunta invisible en el modelo educativo centrado en el docente: “¿con quién aprender?”

Como entorno de aprendizaje, internet es un entorno abierto –demasiado abierto para muchos- de aprendizaje donde su principal valor –junto al acceso a recursos- es la posibilidad de construir redes entre sujetos para aprender y trabajar con metas compartidas. Si hablamos de aprendizaje, la estructura de estas redes puede facilitar o potenciar los procesos de aprendizaje y la generación de conocimiento. No obstante, internet como entorno de interacción social no es uniforme. Hay que tener en cuenta que este entorno social en red, que configura internet, coexisten distintas formas de comunicación, las tradicionales, pero también formas emergentes e inéditas de comunicación y coordinación humana. Este panorama educativo en red implica, por lo menos, tres retos sociales para repensar con quien aprender: 1. Entender internet como ambiente social de aprendizaje más que como un entorno tecnológico. 2. Reconocer qué nuevo rol de aprendizaje se cumple dentro de la estructura social en red, y 3. identificar las transformaciones digitales de las organizaciones educativas en torno a la cultura educativa.

Para mostrar la reconfiguración social potencial se presentan una serie de casos que pueden dar una idea de la amplitud de perfiles y experiencias que ahora existen en la red. No obstante esta selección es parcial, y queda por desarrollar una serie patrones recurrentes en el trabajo en red y una serie de perfiles que abarcan, como se señala en el Peeragogy Handbook (Rheingold et al. 2015) roles potenciales en el proceso de aprendizaje entre pares tanto en los espacios digitales como físicos: el co-líder, co-director del equipo, editor, autor, procesador de contenido, revisor, presentador, comunicador, diseñador, curador, creativo, traductor, estratega, gerente de proyecto, coordinador, asistente, participante, mediador, moderador, facilitador, etc. A pesar de la variedad de nombres que pueda recibir “¿con quién aprender?” hay que reconocer que estamos frente a nuevos roles educativos. Todos estos roles tienen un denominador común: existen porque las condiciones de aprendizaje se han abierto más allá de la docencia y se emplea la red para crear otras formas de relación.

Por tanto, la aportación en la mesa redonda consiste en enfocar internet como un entorno educativo en red donde, además de la enseñanza, caben otros flujos de comunicación y formas de aprendizaje. Esto empuja a redefinir nuestra otredad al momento de pensar la pedagogía y el aprendizaje en distintas disciplinas y un desafío para la práctica de las Humanidades Digitales.

La urgencia de políticas de incorporación de TIC pertinentes en el área de humanidades: estudio de caso, Universidad Veracruzana

Ana Teresa Morales Rodríguez

Alberto Ramírez Martinell

La Sociedad de la Información y el Conocimiento (SIC), es un paradigma postindustrial que asume como materia prima la información, donde se asevera que la productividad está basada en la generación de conocimiento (Castells, 2001). En esto, las TIC son un elemento fundamental que provoca alteraciones, rupturas y rompe esquemas en diversos ámbitos (Brunner, 2005). Cambian los modos de producción (Gibbons, 2001), las formas de socializar son distintas y para cada disciplina hay transformaciones específicas (Becher 2001; Trowler, 2012). Lo que hace necesario que los futuros profesionales sean capaces de incorporar las TIC en sus campos de conocimiento.

Es por eso que las Instituciones de Educación Superior (IES) se han ocupado en incorporarlas en el contexto universitario y ponerlas a disposición de la comunidad universitaria. Entonces, se invierte en tecnología y proyectos de incorporación de TIC para que los egresados sepan usarlas. Pero ¿qué tan pertinentes y efectivos son estos proyectos y políticas?. En el caso de la Universidad Veracruzana (UV), los esfuerzos por incorporar las TIC, se hacen visibles en los planes de desarrollo que los rectores han planteado desde finales de los años 90 (PETIC, 2012), a partir de los cuales se han dictado las estrategias y líneas de acción para la incorporación de las TIC, entre las que destacan: la dotación de infraestructura (equipamiento, conectividad y servicios tecnológicos) para cada una de sus entidades académicas (PETIC, 2012) (las cuales se encuentran distribuidas en cinco regiones geográficas distintas); la implementación de programas de formación en TIC para el profesorado; la puesta en marcha de proyectos como AULA, en el que se capacita a profesores universitarios para el uso de TIC tanto para el diseño de sus clases como para su ejecución; el establecimiento de un marco común de computación para todas las licenciaturas de la universidad; la ampliación de la oferta de servicios de TIC institucionales como iTunes UV, la biblioteca virtual, el sistema de información distribuida Eminus creado al interior de la universidad, entre otros. Dado este contexto, es importante reflexionar que la incorporación de las TIC, implica no solo la dotación de infraestructura, sino que es necesario contar con habilidades digitales para usarlos recursos ya que se desconoce si se está garantizando que los estudiantes de las distintas disciplinas sean capaces de usar las TIC en sus actividades profesionales. Al analizar las políticas de incorporación de las TIC hemos observado que éstas han sido implementadas de manera homogénea (Morales, Ramírez y Excelente, 2016), aún cuando el sistema superior está fragmentado de acuerdo a las diversas disciplinas que convergen al interior.

Implementar políticas de incorporación de TIC de manera homogénea, no permite considerar las necesidades propias de cada disciplina, lo que evita una incorporación pertinente y por lo tanto ineficaz. Es por eso que en el proyecto de Brecha Digital, se ha diseñado una metodología para el análisis de los saberes digitales que requieren los egresados de distintas disciplinas (Casillas y Ramírez 2015). En esta se realizan discusiones colegiadas con las comunidades de académicos de determinada disciplina y se consensan cuáles son los saberes informáticos, informacionales, los requerimientos de comunicación, los elementos de ciudadanía digital y el software especializado, propio de cada disciplina. Esta información sirve como insumo para la actualización de programas de estudio de las carreras en cuestión, y de esta manera la incorporación se plantea que los expertos en cada disciplina (los profesores) puedan definir los saberes digitales mínimos que un egresado debe tener; esos que le permitan apropiarse de tecnologías que puedan usar en apoyo a su disciplina. Así mismo se han llevado a cabo análisis desde diferentes aristas, en los que encontramos que en el área de humanidades hay rechazo al uso de las TIC por parte de algunos profesores, la visión acerca de las posibilidades que éstas les brindan, aún son limitadas. Las pocas innovaciones que hay en el área de humanidades digitales, son escasas y provienen de profesores protecnológicos (Darín, 2015), y a pesar de que éstas resultan enriquecedoras, es necesario y urgente diseñar estrategias que permitan a la universidad encaminar a las carreras de humanidades hacia el uso de las TIC y dejar de tener solo casos de éxito aislados. Es preciso que el análisis y discusión de cuáles son los saberes digitales que deben desarrollarse en los estudiantes de manera diferida en su formación profesional, se plantee a nivel institucional, ya que urge desarrollar en los estudiantes, habilidades que les permitan adaptarse a los cambios tecnológicos, pues de no hacerlo estarían dejando ir una de las grandes oportunidades que ofrece el mundo moderno y en la pasividad se quedarán en el papel de espectadores. Es necesario poder dar a los estudiantes de humanidades (y de todas las áreas), un capital tecnológico que pueda darles ventajas competitivas, hacerlos parte de los cambios y reconfiguraciones actuales, que sean dinámicos, y a su vez congruentes con las necesidades sociales.

Studying Literary Characters and Character Networks

Andrew Piper
andrew.piper@mcgill.ca
McGill University, Canada

Mark Algee-Hewitt
malgeehe@stanford.edu
Stanford University, United States of America

Koustuv Sinha
koustuvsinha@gmail.com
McGill University, Canada

Derek Ruths
derek.ruths@mcgill.ca
McGill University, Canada

Hardik Vala
hardik.vala@mail.mcgill.ca
McGill University, Canada

The study of character has long been one of the central concerns of literary theory. For the Russian formalists, embodied above all in work of Vladimir Propp (1968), character was primarily a "type," one that served different narrative functions ("the hero is married and ascends the throne"). For poststructuralists that came in the wake of Propp, character was nothing more than a rhetorical "effect," one more example of the referential phallacy of naïve readers (Barthes 1970; Culler 2002). Subsequent studies attempted to account for this dual nature of characters, the they are both rhetorical devices and also constrained by real-world references, the requirements of being human that constrain what characters can do and say (Phelan 1989; Jannidis 2004; Frow 2014). More recent research has begun to emphasize the affective or identificatory role that characters play for readers (Lynch 1998; Brewer 2011). According to this view, characters are the media through which readers come to terms with new kinds of social experience. Drawing on the field of cognitive science, other work by Zunshine (2006) and Vermeule (2011) has argued in a less historical vein that characters are useful tools through which to model "theories of mind," means for learning about and hypothetically experiencing human cognition.

It is within this context that our three papers situate themselves in order to understand the ways in which computation impacts the study of literature. Each is fundamentally concerned with how the increased volume of information surrounding characters impacts our understanding of the idea of character – whether it is examining several thousand plays in which characters appear, several thousand interactions between characters, or the millions of words surrounding characters' appearance on the page. Characters are fundamentally social in literature and these computational methods are designed to better understand that sociability.

Mark Algee-Hewitt's paper concerns itself with the study of social networks in 3,900 plays across four-centuries. It asks how the morphology of the social networks represented on stage represent (or resist) both the politics and aesthetics of a period and, more importantly, how those social networks evolve over time? In his paper, he will move beyond the network analysis of a single play by examining the network structure of a large corpus of English dramas written and performed between 1500 and 1920. By applying a series of summary statistics drawn from the field of social network analysis to the individual plays, he is able able to trace the history of dramatic representations of the social sphere and shed new light on the evolution of both the protagonist and the periphery in modern drama.

Khoustiv Sinha, Andrew Piper, Derek Ruth's paper takes a step back to ask the even more fundamental question: what *is* an interaction? Before we move to the extraction and mapping of social networks in fiction, we first need to study how readers understand the very idea of "interaction." In this project, he examines reader annotations across a data set of over 1,000 social interactions drawn from popular contemporary fiction and non-fiction. In doing so, he addresses not only the level of agreement between readers, but also the types of interactions that produce more or less agreement among readers. What are the qualities of social relationships in literature that generate more ambiguity among readers and what does that have to tell us about the social investments of literary texts?

Finally, Andrew Piper and Hardik Vala's paper introduces a new tool that identifies 28 different features aligned with practices of characterization. These features range across a variety of different categories, from positionality (the character's agency), modality (behavior), descriptiveness, to social categories like proximity to other characters or the distribution of character counts in a text As they will show, this tool can allow us to derive novel insights about the history of character development in literary texts.

Distributed Character: Quantitative Models of the English Stage, 1500–1920

Mark Algee-Hewitt

The use of network graphs to represent social networks of characters within novels or plays has played an important role in quantitative textual analysis (see for example Agwar et al 2012, Bingenheimer et al 2011, Elson et al 2010). In this paper, I move beyond the network as a visual object, and instead, draw upon the quantitative metrics of the graph to explore the large-scale changes to the structure of English drama across four hundred years. What can the overall structure of the play can tell us both about the aesthetics of literary production of a given period, and what we can learn about the play by disaggregating the morphology from both the stagecraft and the language that, until now, have made up the two poles of dramatic criticism?

The power of networks lie in their precise, mathematic, description of a set of relationships that can be quantified, measured and aggregated in ways that are unavailable to the reader of a text. Yet, most work has been focused on the use of single networks to describe single plays. For example, in his work on character networks in *Hamlet*, Franco Moretti turns quickly from the quantitative network analysis to the qualitative approach to the plot: "I soon realized that the machine-gathering of the data, essential to large-scale quantification, was not yet a realistic possibility [...] So, from its very first section, the essay drifted from quantification to the qualitative analysis of plot" (Moretti 2011). In this paper, I introduce an automated, rule-based, parsing of the 3439 English plays in the Chadwyck Healy drama corpus in order to perform the kind of large-scale quantitative analysis that Moretti gestures towards, but is unable to realize. The algorithm uses the existing XML markup in the corpus in order to extract speeches and assign them to characters as speakers and recipients, resolving co-references to character abbreviations (this is similar to the automated method employed by Trilcke et al. 2016, although the summary statistics that I extract are quite different).

Drawing on this tagged corpus, I create a social network for each drama and extract a series of summary features based on both the eigenvector and betweenness centralities of each play. The first summary statistic that I calculate is the Gini Coefficient of the eigenvector centrality. Originally designed to measure income inequality within an economic system, the Gini Coefficient is a single number between 0 and 1 that indicates how evenly a set of resources (wealth, income or, in this case, centrality) is distributed across a population (here, of characters. In the Gini coefficient measurement in the corpus at large (Figure 1), there is a clear historical pattern being played out. Over time, between 1550 and 1900, the Gini coefficients of the plays exhibits a clear downward trend, from plays with a small core and a large, non-central periphery in the early century, to plays with a relatively large core and a small periphery in the eighteenth and nineteenth centuries and a large discontinuity between 1650 and 1700. What this metric seems to indicate, then, is the disappearance of the periphery of the English drama over time. Rather than suggesting significant structural changes to the core of the play, the largest influence in the Gini coefficient is the presence of a large periphery, whose members rarely speak (and more importantly, rarely interact with the center of power) and who therefore bring down the Gini coefficient for the entire play. Over time, then, this periphery disappears as casts get smaller and actions take place among an increasingly more tightly knit set of characters. Servants, retainers, guards, acquaintances and messengers, so important during the early modern period, disappear with increasing regularity in the later periods, echoing the reduced function such figures had in society itself, as dramas move, following Habermasian logic, from the throne room to the drawing room, becoming personal and intimate, rather than mythic, political and impersonal.

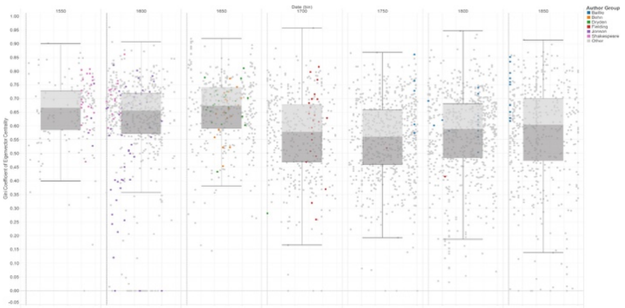


Figure 1. Gini Coefficients of Eigenvector Centrality over time. The corpus is divided into 50 year bins (with the plays in each bin arranged chronologically). Colors indicate selected canonical authors.

The second metric is the percentage of characters in the top quartile of the eigenvector centrality distribution. This measures the size of the core of the play and tells an equally striking and parallel story (Figure 2). Although the relative regularity of the measurement makes it less immediately apparent, there is a constant historical increase in the percentage of characters in the top quartile of eigenvector centrality scores. While the falling Gini Coefficients speak to the disappearance of the periphery, this metric reveals what happens to the remaining core. Rather than follow the same pattern of the early modern period (with few highly central characters), the disappearance of the periphery means that more centrality is allotted between the core characters. This speaks not just to the increasing size of the core, but, more importantly, to the tendency of having plays that feature multiple sub-networks, each with their own protagonist. In a play with a single network, it is easy for one character to dominate it, but in a play whose action is divided between competing communities, each community can have its own central figure. If we can tell the protagonist of an early modern drama by his or her high eigenvector centrality compared to the rest of the cast, then by the seventeenth century the single protagonist has been dispersed between multiple characters who all evidence a high eigenvector centrality, distributing the function of the protagonist (and/or the antagonist) among a growing number of central characters.

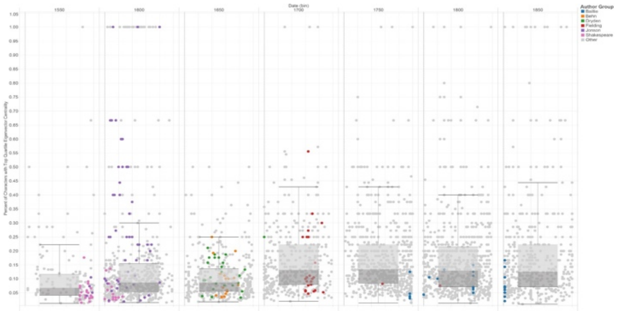


Figure 2. Percentage of characters in the top quartile of the eigenvector centrality distribution in the play.

As opposed to the eigenvector centrality's relationship to the protagonist, betweenness centrality speaks to the mediatedness of the drama. That is, if a high betweenness centrality indicates a character that mediates other character's interactions (such that they have to pass through her), then the scaled maximum betweenness centrality of a dramatic network overall, which measures the relatively importance of bridging characters, indicates the extent to which this mediating function is important to the drama as a whole. At the level of the corpus, the normalized maximum betweenness centrality, the relative importance of the bridging character, decreases across the century, very quickly from 1590 to 1640, and then more slowly across the remaining two and a half centuries: the average maximum betweenness centrality in a play drops by over 750 across just the sixteenth century. Again, the largest discontinuity lies between 1650 and 1700: there is a clearly a lasting effect on the structure of dramatic networks from the puritan shuttering of the theaters during the interregnum. The English drama that returns during the restoration is evidently not the same as the English stage before Cromwell.

Understanding Reader-Identified Social Interactions in Literature

Koustuv Sinha, Andrew Piper, Derek Ruths

Social network analysis begins with the primacy of character as its object of study. In this, it fits within an aready well-established area of inquiry within literary theory, one whose formal study extends back until at least the early twentieth century if not earlier (Propp 1968). Where social network analysis differs from this tradition is through the emphasis on *dynamic interactions* as a key to understanding the narrative function of character. Whether exploring the afterlife of fan fiction, theories of mind, affective identification, or the typologies of character, what all of the pre-computational work on character has in common is an emphasis on understanding character in the singular. Social network analysis argues instead that the meaning of any character is a function of his or her relationships with respect to all of the other characters introduced over the course of a story (Woloch 2009). Character networks offer a way to study not simply the types or themes or affective connections between readers and imaginary people. Rather, they afford us the ability to understand the social imaginings of writers, periods, and genres.

Several initial attempts to introduce social network analysis into the study of literature have already been made. Character networks have been studied within three major European epics to understand their relation to contemporary models of social networks (MacCaron/Kenna 2012); an abridged version of a single well-known literary work (*Alice in Wonderland*) to test differences between interactions and observations on character centrality (Agarwal 2012); nineteenth-century novels to understand the correlation between dialogue and setting (Elson 2010); as a form of narrative generation (Sack 2013); and the genre of classical drama to better understand the notion of tragic conflict (Moretti 2013; Karsdorp et al 2015).

Each of these works has added to our understanding of the relationship between character and literary form in important ways. And yet at the core of each of these studies lies a fundamental assumption about the self-evident nature of an "interaction." Initial attempts to use machine learning to derive interactions on prose texts have shown very poor performance (Agarwal 2012 reports a maximum F1 score of 0.61). What this indicates at least in part is that interactions are highly complex verbal constructions which we cannot easily assume pre-exist our attempts at extracting them.

To counter this problem, we have designed a study to explore reader agreement across a variety of text passages (1,000) drawn from popular contemporary fiction and nonfiction. Rather than begin with a stable set of interaction types, however, our goal is to infer possible classes of interactions and then understand which of these classes generate more ambiguity among readers. We perform this in three phases. In the first phase, we ask coders to identify minimally defined interactions using a standardized web interface (where an interaction consists of two entities and an action linking them). Our goal here is not to pre-define types of interactions as in other studies (Agarwal), but to better understand how readers intuitively understand social interactions between characters. As we have shown in another study, readers indicate very high agreement in identifying character aliases (i.e. determining what is an entity (Vala et al.)). In the second phase, we use unsupervised clustering techniques to identify different interaction "types" based on syntactic and lexical features of the labeled interactions. Third, we then measure reader agreement across these different types. While we want to know overall how well readers agree on defining interactions, we also want to understand if different types of interactions across different types of writing (fiction/non-fiction) illustrate signigicantly higher levels of disagreement. This is a first step in understanding the unique ways literary texts generate social complexity, not simply through the quantity of interactions but also importantly through their qualities.

Emma: A Feature Space for Studying Character

This paper will argue that computation has an important role to play in understanding the nature of characters and the process of what we might generally term characterization – the writerly act of generating animate entities through language. With an estimated 86 characters per novel in the nineteenth century and a conservative estimate of 20,000 novels published during this period in the English language, there are over 1.7 million unique characters that appear in that one century and one language alone. Even if we condition on main characters, we are still looking at several thousand distinct entities. At the same time, there are not only a great number of characters in literature, but there is also a tremendous amount of information surrounding even one primary character. Like other highly frequent textual features such as conjunctions or punctuation, characters are abundant across the pages of individual novels. Personal pronouns alone account for roughly 12% of all tokens, and if one adds in proper names the number of character occurrences is closer to 16% – or one in every six words! Like the abundance of characters, such semiotic abundance surrounding characters poses problems for inherited critical methods. How can we be sure that our claims about "character" are capturing the broad and potentially diverse ways that characters are depicted in novels, this larger mass of fictional beings and what it means to be fictional?

In order to address this question, we have developed a computational tool designed for the study of character. Its aim is to identify 28 different features that relate to qualities that characters may possess. These range across categories like distinctiveness (how distinctive is the main character from other characters within the novel); positionality (how often is the character the agent or object of a sentence or a possessor of some object); centrality (how important is the protagonist relative to other characters in the novel); and modality (what kinds of behaviors and descriptions inform this character's identity, such as cogitation, perception, motion, embodiment and even clothing or dress).

Rather than start with known "types" of character, this tool allows us to implement a more multi-dimensional understanding of character and use that representation to think about the relationships between novels. Prior work on stylistic analysis has not differentiated between various aspects of texts when comparing them to each other. The novel is taken as a unified whole. Our character feature tool allows readers to begin to explore these different sub-domains of a novel, which in our case refers to the language used to construct character. In our presentation, we will discuss the mechanics that underlie the tool, which implements a modified version of BookNLP (Bamman 2014) and the Stanford dependency parser in order to identify words related to character. We will also discuss a case study in which we explore the identity of "introversion" in novels from the nineteenth century to the present. As we will show, the character feature tool allows us to construct not only familiar narratives about the history of the novel – wherein the representation of interiority is strongly gendered around female protagonists – but also novel and nuanced insights about that tradition when we follow these features across a broader swath of time. As we will show, interiority no longer remains the distinctive quality of feminine heroines but is transposed onto a very different generic and gender scene – the male hero of science fiction.

Reconstruction –Representation –Collaboration: Interdisciplinary approaches to changes in contexts of digital (music) editions

Franziska Schloots
franziska.schloots@uni-paderborn.de
Paderborn University, Germany

Bianca Meise
bianca.meise@uni-paderborn.de
Paderborn University, Germany

Peter Stadler
stadler@weber-gesamtausgabe.de
Paderborn University, Germany

Jörg Müller-Lietzkow
jml@mail.uni-paderborn.de
Paderborn University, Germany

Dorothee M. Meister
dorothee.meister@uni-paderborn.de
Paderborn University, Germany

Johannes Kepper
kepper@edirom.de
Paderborn University, Germany

Daniel Röwenstrunk
roewenstrunk@edirom.de
Paderborn University, Germany

Introduction

The field of Digital Humanities is characterized by complex questions and interdisciplinary discussions. That applies not only to the cultural science-based side of the humanities but also to the digital representation of cultural artefacts and, among other things, their processing, analysis, preservation and long-term availability. In order to handle these issues, interdisciplinary collaboration is required to pool and to coordinate expert knowledge and skills.

The Digital Humanities project ZenMEM (Centre of Music – Edition – Media) represents such an interdisciplinary project. Since September 2014, researchers from Paderborn University, the Hochschule für Musik Detmold and Ostwestfalen-Lippe University of Applied Sciences combine their expertise from musicology, various fields of computer sciences (contextual informatics, software engineering, usability engineering and music informatics) and media studies (media education and media economics) to investigate processes of change and new possibilities of the transition from analogue to digital music and media editions.

In this context, digitization can be discussed as distributed through material infrastructures and displayed on computer devices (cf. Huber 1998). But it can also be considered in non-physical dimensions. In this perspective cultures are containing specific structures that build a para-material quality (cf. Schrage 2006).

> "Writing is a material act; textual production in any medium has always been a part and product of particular technologies of inscription and duplication." (Kirschenbaum et al. 2009).

A third perspective can be discovered between these poles on the transition from material and immaterial (cf. Holl 2010, Manovich 2011, McGann 1991). The fourth dimension in this context refers to the fact that

> "Digital Humanities work embraces iterative, in which experiments are run over time and become object to constant revision. Critical design discourse is moving away from a strict problem-solving approach that seeks to find a final answer: Each new design opens up new problems and – productively – creates new questions." (Burdick et al. 2012)

Between these perspectives, we explore interdisciplinary approaches of digital (music) editions. First, the dimension of digital representation and the contents on the surface (cf. Manovich 2001). Here, the traditional Human Computer Interface (HCI) to digital editions is questioned and expanded to incorporate Computer Interfaces, ultimately demanding to "granting access to machines".

The second perspective focusses on the relevance of the user for all types of media (cf. Fiske 2001). In this view, we present a qualitative research study (cf. Denzin 2000; Przyborski & Wohlrab-Sahr 2009; Keuneke 1999) dealing with the changing work processes of musicologists and music editors in the context of digital music editions. These changes have, among other things, lasting impacts on their scientific research, orientation, academic ethos and knowledge building.

The third paper deals with the structural changes in scientific work in context of Digital Humanities. A growing number of interdisciplinary academic projects causes an increasing need for researchers with the key qualification "interdisciplinarity" as well as methods to verify the projects' efficiency and quality (cf. von Kardorff 2000).

The analysis of the complex interactions among these dimensions leads to essential issues of the digitization of (music) editions: The access to perspectives of digital representation, the impacts on scientific work and access to user views and at last a dimension of access to changing interdisciplinary project work within the Digital Humanities.

In case of acceptance, this multiple paper session would be presented by Bianca Meise (media education and qualitative research, Paderborn University), Peter Stadler (musicology, Paderborn University) and Franziska Schloots (media economy and management, Paderborn University).

Digital Editions and the Interface. Granting Access to Machines

Peter Stadler

Johannes Kepper

Daniel Röwenstrunk

The history of Digital Editions and its main driving force, the Text Encoding Initiative (TEI), is a success story. It not only enabled the digital presentation and preservation of scholarly editions but subsequently facilitated further research on this digital material and the development of tools and methods, pushing forward the DH sector as a whole.

The TEI puts a lot of effort into its (prose) Guidelines and the formal specification of the schema(s) – it ultimately developed a meta language (ODD = One Document Does it all) for the definition and documentation of TEI schemas and customizations, leading to a well-documented, highly flexible interchange format. "This is standardization by not saying 'Do what I do' but instead by saying 'Do what you need to do but tell me about it in a language I understand'." (Cummings 2013)

The Music Encoding Initiative (MEI) adopts these principles, empowering it to encode a wide variety of musical styles and genres through its modular approach. Yet, a long-standing issue with the resulting XML documents is the lack of interoperability. True interoperability would allow the 'blind' reuse of files within one's own processing chain (cf. Bauman 2011). But due to the TEI's and MEI's flexibility of encoding even the (assumed) simplest operations can hardly be processed blindly, e.g. the extraction of a plain text version (aka the "throwing away of angle brackets") or the extraction of a single voice from a score.

Generally speaking, digital editions are potentially multifunctional and enable multiple views on the text (cf. Sahle 2013; Pierazzo 2016), yet this potential can hardly be activated from the outside but has to be revealed by a standardized endpoint or interface. Most commonly, the only (public) interface of a digital edition is the HTML version accessible online with a (Javascript enabled) browser. Different views are created from one TEI source file and prepared for a human reader, e.g. a “Semantic Edition” or a “Philological Version” (for example, see the respective tabs of the letters published on Burckhardt Source). While this interface and its functionality is adequate and pleasing for a human agent, it’s more or less useless for a machine agent which would have to grab the web pages and parse idiosyncratic flavours of (X)HTML(5). Only a few digital editions offer the direct download of TEI encoded files and even less offer a dedicated (and well documented) interface for machines.

Within the ZenMEM and ViFE research infrastructures we are trying to remedy this shortcoming of digital editions by developing dedicated APIs for our projects. We believe that this has at least two advantages:

1. intrinsic: better documentation of our own work
2. extrinsic: facilitated reuse of our work by others

These project APIs are developed and documented using the Swagger framework and the resulting configuration files can be found on Github. Along with those project specific APIs we strive to develop a meta API with a core set of functions which are to be supported by all projects. There is an ongoing discussion on what these core functions are – especially since we are dealing with various materials, from music to text encoding, from placeographies to bibliographies, from sketches to prints. On the other hand, we are confident to come up with a set of generic functions that could then easily be adopted by other projects. With these interfaces to digital editions we will finally leave the traditional resource based approach to digital editions behind and move on to a new functional, truly dynamic approach.

A changing paradigm? Implications of digitization on musicologists work on editions

Bianca Meise

Digital Humanities are focusing on digital data. Therefore, issues on modelling, representation, analysis and annotations are crucial dimensions of research like processing and archival storage. But digital data and the procedures quoted before are used by editors and have impacts on their scientific work, too (cf. Edwards 2012). This contribution is focused on an access to the editors as special user and producer group in the process of digitization through a qualitative empirical study. In this study, it is neither the editor nor the data alone, rather, it is the act of editing, analyzing, representing and annotating that gives insights into the relation between media, material and subjects that allow deep insights into the transformational impacts of digitization. First, I discuss the changes of the scientific working process from analogue to digital. Secondly, the challenges and potentials of this change of paradigm will be discussed. At last, as a result of these findings, the future inquiries of digitization of music editions, changes of work processes and not least the education and varying access to knowledge work will be pointed out.

Digital music editions have a lot of potential for editors, scientists and recipients (cf. Veit). Thus, the digital availability accelerates and arranges the editorial scientific process. Furthermore, the different representations of the digital offer a great transparency and confirmability. In this perspective, the digitization of music editions enriches the work of musicologists (ibid.). A lot of issues are discussed in context of digital humanities refer to the digital representation or the transformation of cultural artefacts, their further analysis and processing. These subjects are considered on data. The work, examination and handling of the digital and the influences of the knowledge evolving within is discussed less. In this contribution, the formal layer (cf. Kirschenbaum 2008) or cultural layer (cf. Manovich 2001) and moreover the performative interactions between the material and the subjects will be considered (cf. Drucker 2013). Issues like the challenges to adopt digital techniques, to prepare digital sources, to represent and analyze them are important, too. The examination of the various practices of adoption, handling and orientation offer deep insights into the possibilities of digitization and its relevance for scientific work. This study allows a rare attendant insight in the transformation of a media paradigm in music philologists’ work.

The editions contain different types of "texts" like several notation formats, facsimiles, born digital annotations, text and audio files. Before digitization there are analogue procedures of searching, collecting, arranging and reviewing. Since the digitization of music editions, the shift is not only digital. Moreover, it’s both digital and analogue, material and corresponding practices, that construct the editions. It is the operation, the handling of the musicologist with cultural, material and immaterial artefacts. To develop the importance of media and the corresponding methods in digital music editions, its necessary to explore these various interactions. Like the radical contextualization of the cultural studies pointed out that object and subject, media technology and context (cf. Winter 2010) continuously affecting each other and be interwoven.

Based on the qualitative research perspectives of the Grounded Theory Methodology (cf. Strauss & Corbin 1996; Denzin 2000, Przyborski & Wohlrab-Sahr 2009) in this study eight narrative guided interviews with problem-centered parts have conducted. Because of the academic void of the user perspective qualitative research offers great potentials to get essential insights and generate first hypothesis in this field (cf. Flick et al. 2000). The investigation on varying routines or the knowledge of the musicologists even if it's implicit (cf. Polyani 1985) is not simple to transfer in direct questions. In fact, the qualitative research provides many methods to manage this problem (cf. Flick 2002). The guided interview is an inquiry method that follows a structure, but allows a lot of free space for further situated questions and narrative answers (cf. Keuneke 2000). The questions are focused on the routines of the working process of music philologists, their evolving (implicit) knowledge and their biographical reference points to music editions. The qualitative material has been analyzed by a modified coding version of the grounded theory (cf. Strauss & Corbin 1996) like Przyborski and Wohlrab-Sahr (2009) pointed out.

The empirical data and their interpretation documented, that the change in the working process in not only a shift from analogue to digital. Moreover, the hole scientific process is changed and the work on digital music editions have deep impacts on editorial, juridical, organizational, social and not at least educational processes. Digitization changing the scientific work organization, the editorial work and the perspective on editions and the interwoven knowledge.

Interdisciplinary research and knowledge building – Development of recommendations for action for academic projects in the digital humanities

Bianca Meise,

Franziska Schloots

Jörg Müller–Lietzkow

Dorothee M. Meister

The heterogeneous research field of digital humanities is characterised by a high level of connectivity and interdisciplinarity. Not least because of this the digital humanities turned out to be a dynamic field with high velocity of scientific discourses (cf. Gold 2012). In this context, the development of research infrastructure for interconnected and collaborative scientific work (cf. Reichert 2014) as well as the management of interdisciplinary research projects represent particular challenges. Whereas especially the economically characterised management research (cf. Bea, Scheurer & Hesselmann 2011/ Steinmann & Schreyögg 2005) deals with project management processes in companies, network groups or other cooperation forms, project management in academic contexts is hardly found in research literature. One important distinctive feature, especially for academic projects without industrial partners, is that the goal dimension of academic projects fundamentally differs from economic projects. Its focus is not a typical measurable economic gross profit or other definable objectives. The openness for results as an important part of scientific research requires a modified form of project management. However, strategic and systematic project management is hardly implemented in the structure of universities, particularly in cultural science departments and projects. Oftentimes, a rather “muddling through” (Lindblom 1959) can be noticed. This is not a desirable condition because consequences might be increasing transaction costs as well as reduce chances for long-term and stable research collaborations. Furthermore, the special structures in academic contexts must be considered concerning the planning, implementation and organisation of collaborative interdisciplinary projects.

In view of these conditions, the importance of professional project evaluation is evident, especially in interdisciplinary projects involving cultural scientists as it is the case in digital humanities projects. Digital humanities can be seen as an extension of the complexity dimension according to Rinza (1998): A high scientific novelty grade is associated with an increased risk for the project goals and a relatively large project group might increase the dependencies in the work processes because of the close interlocking of the individual departments.

The growing concentration of interdisciplinary academic projects causes an increasing need for scientifically verified evidence of their efficiency and quality (cf. von Kardorff 2000). For this purpose, summative evaluations are used to review the effectiveness and the achievement of defined targets as well as formative evaluations which can contribute to the optimization of an ongoing project and can provide the basis of strategic decisions and changes (cf. ibid.). Within this paper, the procedure will be clarified through a concrete case study from the field of digital humanities.

In the project ZenMEM (Centre of Music – Edition – Media), a formative evaluation in the ongoing project was conducted so its results could be used to develop recommendations for action for the further process.

In ZenMEM, researchers from Paderborn University, the Hochschule für Musik Detmold and Ostwestfalen-Lippe University of Applied Sciences investigate processes of change and new possibilities of the transition from analogue to digital music and media editions since September 2014.

The project is initially promoted for three years by the German Federal Ministry of Education and Research (BMBF) and combines experiences and expertise as well as concepts and methods from musicology, various fields of computer sciences (contextual informatics, software engineering, usability engineering and music informatics) and media studies (media education and media economics). The focus of their research is on musical and other non-textual objects in context of digital editions. In this sense, the researchers are able to link to their preliminary scientific work as well as international developments like the Edirom project or the standards of Music Encoding Initiative (MEI) and Text Encoding Initiative (TEI) . They participate in its further development and examine innovative interacting and editing functions for the creation of digital music and media editions. In addition to research work, corresponding software tools are developed, technical advice and coordination is provided to external projects and training activities in form of workshops, courses and lectures are conducted and further developed. These steps are being accompanied by qualitative and quantitative user studies which take the entire creation process of digital editions into consideration. The results flow back into research and development work within the ZenMEM center. This short project description gives an impression of the complexity of the tasks in this joint project and the amount of heterogeneous research questions the cooperation partners are dealing with.

To conduct a formative evaluation for the ZenMEM project in order to develop recommendations for action for the further project progression, a qualitative half-standardized written survey was chosen as research method. For this purpose, a questionnaire with 18 questions was given to all project members. These questions were relating to previous experiences with working in interdisciplinary projects, special challenges within the project, specific work packages and personal goals. In the last section, the researchers were asked for valuation of ongoing processes and constructive suggestions for improvement. In addition, a shorter questionnaire was sent to the project leaders. Thus, the perspectives of different project status groups could be surveyed.

For the analysis of the results, amongst other things, a categorization based on the concept of SWOT analysis was performed. Developed at Harvard Business School in the 1960s for the strategic planning in companies, the SWOT analysis is a versatile tool (cf. Schawel & Billing 2009) describing the strengths, weaknesses, opportunities and threats of any company or project which constitute a basic structure for developing strategic recommendations for actions (cf. Kotler, Berger & Bickhoff 2016). To derive concrete statements from these four dimensions, the categorization had to be refined. Regarding the evaluation method of coding with Grounded Theory (cf. Przyborski & Wohlrab-Sahr 2009) a differentiated categorization within the root categories was performed.

In this context, it is important to note that Grounded Theory is a methodology as well as an evaluation method (cf. Strübing 2004). The paper at hand focuses on Grounded Theory as an evaluation method to collect phenomena, condense them into concepts, identify categories and uncover connections between them. Following, recommendations for action could be worked out from the combinations of single phenomena in order to optimize processes within the project. Furthermore, indications could be provided to identify which actions should be focused and which should be avoided (cf. Pepels 2005). Individual competences of the project participants were also taken into consideration. Among other things, it became clear that interdisciplinarity in such a project can be considered as both opportunity and challenge, especially regarding the different preconditions, approaches and methods. People not only hold various paradigms and epistemological interests but also act within different organizations and social systems – nearly all university departments and courses are oriented mono-disciplinary (cf. Dressel et al. 2014). Stringent guidance is needed to enable successful interdisciplinary research. In the ZenMEM project, superiority of the subject disciplines was at least a little bit softened and reflected by mutual work observations as well as agreements on certain terminology so that a basal equality of the participating research partners could be established and constantly evolved.

While in this case, systematic knowledge building took place directly from within the project, it can be discussed how the mediation of interdisciplinary competences could be implemented a lot earlier. As already indicated, interdisciplinary collaborations in academic contexts are of increasing significance and the demand for qualified young researchers with the key qualification interdisciplinarity (cf. Mainzer 2013) grows. More and more universities offer degrees in digital humanities or related fields. However, corresponding curricula show very different weightings of cultural science-based or computer science-based contents (cf. Schubert 2015). A closer inspection of digital humanities projects and a comparison between project experience and curricula might be helpful to develop degree courses so that interdisciplinary competences and shared knowledge can be gained.

Bibliography

Bauman, S. (2011): “Interchange vs. Interoperability.” In: Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7 http://dx.doi.org/10.4242/BalisageVol7.Bauman01 [01.11.2016]

Bea, F. X., Scheurer, S. & Hesselmann, S. (2011): Projektmanagement. Konstanz: UVK Verlagsgesellschaft.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2012): DIGITAL HUMANITIES. MIT Press.

Cummings, J. (2013): “ODDly Pragmatic: Documenting encoding practices in Digital Humanities projects.” https://blogs.it.ox.ac.uk/jamesc/2013/09/21/oddly-pragmatic/ [01.11.2016]

Denzin, N. K. (2000); Symbolischer Interaktionismus. In: Flick, Uwe; Kardorff, E.; Steinke, I. (Hrsg.): Qualitative Forschung. Ein Handbuch. Hamburg: Rowohlt.

Dressel, G., Heimerl, K., Berger, W. & Winiwarter, V. (2014). Interdisziplinäres und transdisziplinäres Forschen organisieren. In: Dressel, G., Berger, W., Heimerl, K. & Winiwarter, V. (Hrsg.) Interdisziplinär und transdisziplinär forschen. Praktiken und Methoden. Bielefeld: Transcript S. 207-212.

Drucker, J. (2013): Performative Materiality and Theoretical Approaches to Interface. In: Digital Humanities Quartely, Vol. 7 (No.1). Online verfügbar: http://www. Digitalhumanities. org/dhq/vol/7/1/000143/000143.html. [01.11.2016]

Edwards, C. (2012): The Digital Humanities and Its Users. In: Gold, M. K. Debates in the Humanities. See also http://dhdebates.gc.cuny.edu/debates/text/31. [01.11.2016]

Fiske, J. (2001): Die populäre Ökonomie. In: Winter, R./ Mikos, L. (Hrsg.): Die Fabrikation des Populären. Der John Fiske-Reader. Bielefeld: transcript, S. 111-137.

Flick, U. (2002). Qualitative Sozialforschung. Eine Einführung. 6. überarb. und erweiterte Aufl. Hamburg: Rowohlt.

Flick, U., Kardorff, E. von, Steinke I. (Hrsg.). (2000). Qualitative Forschung. Ein Handbuch. Reinbeck: Rowohlt.

Gold, M. K. (2012): The Digital Humanities Moment. In: Gold, M. K. (Hrg.) (2012): Debates in the Digital Humanities. Minneapolis: University of Minnesota Press, S. IX-XVI.

Holl, U. (2010): Materialität/ Immaterialität. In: ZfM Zeitschrift für Medienwissenschaft (2), Berlin: Akademie Verlag, 10-14. http://www.zfmedienwissenschaft.de/heft/text/materialit%C3%A4t-immaterialit%C3%A4t [01.11.2016]

Huber, H. D. (1998): Materialität und Immaterialität der Netzkunst. In: kritische berichte, Zeitschrift für Kunst- und Kulturwissenschaften, Sonderheft Netzkunst, Jg. 26, 1998, Heft 1, S.39-53.

von Kardorff, E. (2000): Qualitative Evaluationsforschung. In: Flick, U., von Kardorff E. & Steinke, I. (Hrsg.) (2000): Qualitative Forschung. Ein Handbuch. Reinbek bei Hamburg: Rowohlt, S.238-250.

Keuneke, S. (2005). Qualitatives Interview. In: Mikos, L. & Wegener, C. (Hrsg.) Qualitative Medienforschung. Ein Handbuch. Konstanz: UVK, S. 254-267.

Kirschenbaum, M. (2008): Mechanisms: New Media and the Forensic Imagination. Cambridge: MIT University Press.

Kirschenbaum, M. G. et al. (2009): Digital Materiality: Preserving Access to Computers as Complete Environments. http://escholarship.org/uc/item/7d3465vg [01.11.2016]

Kotler, P., Berger, R. & Bickhoff, N. (2016): The Quintessence of Strategic Management. What You Really Need to Know to Survive in Business. Second Edition. Berlin/ Heidelberg: Springer.

Lindblom, C. E. (1959): The Science of „Muddling-Through". Public Administration Review, 19 (2), 79-88.

Mainzer, K. (2013): Interdisziplinarität und Schlüsselqualifikationen in der globalen Wissens-gesellschaft. In: Jungert, M. et al. (Hrsg.). Interdisziplinarität. Theorie, Praxis, Probleme. Darmstadt: WBG, S. VI-VIII.

Manovich, L. (2001). Language of New Media. Cambridge: MIT Press.

Pepels, W. (2005): Grundlagen der Unternehmensführung. Strategie – Stellgrößen – Erfolgsfaktoren – Implementierung. München: Oldenbourg Wissenschaftsverlag.

Pierazzo, E. (2016): Digital Scholarly Editing: Theories, Models and Methods. Routledge.

Polanyi, M. (1985). Implizites Wissen. Frankfurt: Suhrkamp.

Przyborski, A & Wohlrab-Sahr, M. (2009): Qualitative Sozialforschung. Ein Arbeitsbuch. München, Oldenbourg.

Reichert, R. (2014): Digital Humanities. In: Schröter, J. (Hrg.) (2014): Handbuch Medienwissenschaft. Weimar: Verlag J.B. Metzler, S. 511-515.

Rinza, P. (1998): Projektmanagement. Planung. Überwachung und Steuerung von technischen und nichttechnischen Vorhaben. 4. neubearbeitete Auflage. Berlin/Heidelberg: Springer.

Sahle, P. (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Norderstedt.

Schawel, C. & Billing, F. (2009): Top 100 Management Tools. Das wichtigste Buch eines Managers. 2. Auflage. Wiesbaden: Gabler/ GWV Fachverlage.

Schrage, D. (2006): Kultur als Materialität oder Material – Diskurstheorie oder Diskursanalyse?" In: Rehberg (Hg.): Soziale Ungleichheit – Kulturelle Unterschiede (Bd. 2), S. 1806-1813. http://blog.dominikschrage.de/wp-content/uploads/2014/04/material.pdf [01.11.2016]

Schubert, Z. (Red.) (2015): Digital Humanities als Beruf. Fortschritte auf dem Weg zu einem Curriculum. Akten der DHd-Arbeitsgruppe „Referenzcurriculum Digital Humanities". Vorgelegt auf der Jahrestagung der DHd 2015 vom 24. Bis 27. Februar 2015 in Graz. Online abrufbar unter https://www.digitalhumanities.tu-darmstadt.de/fileadmin/dhdarmstadt/materials/Digital_Humanities_als_Beruf_-_Stand_2015.pdf. [27.10.2016]

Steinmann, H., Schreyögg, G. (2005): Management. Grundlagen der Unternehmensführung. Konzepte, Funktionen, Fallstudien. Wiesbaden: Gabler.

Strübing, J. (2004): Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung des Verfahrens der empirisch begründeten Theoriebildung. Wiesbaden: VS.

Straus, A. & Corbin, J. (1994): Grounded Theory Methodology. An Overview. In: Denzin, N. K. (Hrsg.).: Handbook of Qualitative Research. London, New York: Sage, S. 279.

Suber, P. (2012): Open Access. MIT Press. http://bit.ly/oa-book [01.11.2016].

Winter, R. (2010): Handlungsmächtigkeit und technologische Lebensformen. Cultural Studies, digitale Medien und die Demokratisierung der Lebensverhältnisse. In: Pietraß, M. & Funiok, R. (Hrsg.) Mensch und Medien. Philosophische und sozialwissenschaftliche Perspektiven. Wiesbaden: VS, S. 139-157

Refining our Concept of 'Access' for Digital Scholarly Editions: A DiXiT Panel on Accessibility, Usability, Pedagogy, Collaboration, Community and Diversity

Anna-Maria Sichani
anna-maria.sichani@huygens.knaw.nl
Huygens ING - KNAW, The Netherlands

Wout Dillen
wout.dillen@hb.se
Swedish School of Library and Information Science
University of Borås, Sweden

Melisa Ariel Martinez
merisa.martinez@hb.se
Swedish School of Library and Information Science
University of Borås, Sweden

Aodhán Kelly
aodhan.kelly@uantwerpen.be
Center for Manuscript Genetics

University of Antwerp, Belgium

Federico Caria
federico.caria@uniroma1.it
Sapienza Università di Roma, Italy

Elli Bleeker
elli.bleeker@uantwerpen.be
Center for Manuscript Genetics
University of Antwerp, Belgium

Introduction

Access, in all its iterations, continues to shape the discourse of digital scholarly editing as the field grapples with new models and methods. Our proposed 90 minute panel will frame a discussion around a broader definition of the concept in relation to the field of digital textual scholarship, by critically reflecting on its meaning for Digital Scholarly Editions (DSEs) and theorizing how the term relates to issues of accessibility, usability, pedagogy, collaboration, community,and diversity. Each of the fellows will make short '7-14-28' presentations (seven minutes for 14 slides in 28-pt font), identifying results and questions arising from our research over the last three years and leaving 48 minutes for discussion. Refining our concept of access signifies a valuable contribution to the field: while 'accessibility' is a highly-cited term in digital editing, its use generally refers to making data (Sahle 2014) and source materials (Martens 1995: 222) more openly available rather than to making data more understandable to different types of users, including users with disabilities. Similarly, discussions regarding different user needs typically refer to those with a non-academic background (e.g. Apollon et al. 2014: 93; Pierazzo 2015: 151), rather than users with (in)visible disabilities.

Accessibility and Usability

The digital medium gives the DSE the capacity to be more accessible than its predecessor. Automated analysis and the processing of algorithms allow for the development of a host of tools useful to people with disabilities. Existing tools like screen magnification software or Text To Speech software , for example, already help people with visual impairment to better absorb and navigate the edition's contents. But such ready-made user-dependent solutions only scratch the surface of the ways in which we can make our editions more accessible to people with both visible and invisible disabilities. When designing a web interface for DSEs, current accessibility standards in web design are rarely taken into account — if indeed at all. The mere fact that two major points of reference in our field (Sahle 2014; Franzini 2016) do not mention accessibility in their respective lists of criteria already implies that standards such as @alt texts for links and images, consistent use of header tags, legibility of fonts, attentive use of colors and contrast, etc. are not sufficiently acknowledged or adopted. As editors, perhaps our preoccupation with the underlying XML of our editions has lead us to be less conscientious about our use of the transformed HTML. In this respect, we could follow the example set by the World Wide Web Consortium, which has made a considerable effort in encouraging an increased coordination between the interrelated concerns of accessibility, usability, and inclusion in web design (W3C 2016).

Usability is inextricably linked with design approaches and practices, both in terms of functions and aesthetics. Ruecker et al. (2011: 13) argue that aesthetic design promotes the perceived usability as well as the perceived value and legitimacy of digital cultural heritage materials. Editing projects can develop these skills internally as part of a change in pedagogy, or, in order to strengthen a community of practice and knowledge transfer, can develop them in collaboration with other disciplines, libraries, and private sector partners. Measuring usability and access can also be gauged through opportunities for reuse. This would involve determining the extent to which edition data is made available for open access and what type of licensing information for potential re-use is communicated to the user. Mapping the W3C's standards onto the practice of scholarly editing is not only a good way of supporting the community of people with disabilities, but also a constructive way to confront the digital divide and generally benefit users of DSEs.

Pedagogy and Collaboration

As new publication models emerge in this developing field of scholarly research, practitioners are continually expected to readjust their skillsets. Pierazzo (2016) noted that the role of the editor has changed significantly in the digital medium, growing from advanced awareness of classical theory to the 'super-editor' model with added requirements of technical skills including understanding of front and back-end web design, image processing, XML, specialized scripting languages, etc. In an environment where these skills are not only desired but also expected, aspiring editors spend significant time and money on acquiring them through workshops, courses, and prolonged research stays, in which specific projects provide opportunities for in-depth training as well as valuable networking opportunities.

Acknowledging our place of privilege in this debate as the recipients of a European research grant, we would argue that while our network's practice of sponsoring conference and workshop attendance for non-DiXiT early career researchers is a necessary first step, it may not be enough: we also need to rethink the way these courses are offered, and to to develop best practice documents for training new editors. To accomplish this, we need to ask ourselves whether we want to focus on a broad and comprehensive skillset with an overall understanding of concepts and principles, or accept a specialisation ethos that invites more collaborative work. Our own training has primarily involved literary or historical materials using the latin alphabet and reflecting dominant editorial schools, traditions, and scholarly disciplines. Indeed, the majority of the training material produced for digital editing (workshops, seminars, books, guidelines, etc.) is produced in English or in western-European languages, thereby excluding scholars from smaller communities from fruitful engagement with the field. How do we actively promote a more inclusive approach both to the content of editions and to the training of future editors? Can we adopt a vantage point that justifies the inclusion of and training in a variety of disciplines, without diminishing the value of highly specialized knowledge?

Community and Diversity

Overwhelmingly, DSEs have focused on the documents and narratives of western-European and North American males (for example, the catalogues of Sahle (2016) and Franzini (2016) predominately feature this demographic). We need to be critical in asking ourselves hard questions about our assumed audience, gender-equity and social justice, and which narratives we are gatekeeping by choosing these texts. Given that our editions increasingly do not reflect the gender array of our practitioners, a reorientation toward underrepresented voices is called for. This is modelled by digital libraries and archives, which offer significant collections about women's history, LGBTQIA culture, people of color, indigenous peoples, and people with disabilities. It would behoove digital editors to follow the example set by our collaborators and seek out opportunities to expand the canon, attributing the same care and attention to texts produced by these groups.

Finally, along with thinking of ways to expand the canon of scholarly editions, we also need to reflect on new ways to diversify the community of scholarly editors — which, like its subject, could also be characterised as a predominantly white, Western-European or North American male community. Bordalejo (2016) presented a similar argument about demographics in DH that could easily be extrapolated to our subfield. In a recent paper Robinson (2016) called for a reconsideration of the role of editors and scholars by taking a more social approach whereby these academics should become 'key participants in, and enablers of, communities' rather than leaders of more exclusive collaborations. This encourages us to reflect thoughtfully about how digital scholarly editing is conceived and performed as an elite activity, accessible mostly to researchers and communities with sufficient financial, infrastructural, and societal means to support them. If we recognize this, how can we encourage a more inclusive approach?

On the Panel

The Digital Scholarly Editions Initial Training Network (DiXiT) is a Marie Skłodowska-Curie EU-Funded 7th Framework Programme. During the grant period (2013-2017), twelve Early Stage Research Fellows and five Experienced Research Fellows engage with questions and tensions surrounding the evolving theory and practices of digital scholarly editing. As our projects draw to a close we are reflecting critically on how we have examined and contributed to the changing nature of digital textual scholarship. With access being such a pertinent issue to the field of digital textual scholarship, we hope to stimulate a lively and productive conversation with the audience around these interrelated themes.

Bibliography

Bordalejo, B. (2016) 'Diversity in Digital Humanities.' Paper presented at *DHBenelux 2016*. 9-10 June 2016: Centre Virtuel de la Connaissance sur l'Europe (CVCE) and the University of Luxembourg, Luxemburg.

Franzini, G. (2016) 'Catalogue of Digital Editions.' *Github*. Available online on https://github.com/gfranzini/digEds_cat/wiki (accessed on 1 November 2016; last updated 12 June 2016).

Martens, G.(1995) 'What is a Text? Attempts at Defining a Central Concept in Editorial theory' In: *Contemporary German Editorial Theory*, edited by Hans Walter Gabler, George Bornstein, and Gillian Borland Pierce. Ann Arbor: University of Michigan Press.

Pierazzo, E. (2015) *Digital Scholarly Editing: Theories, Models and Methods*. Surrey: Ashgate Publishing, Ltd.

Pierazzo, E.(2016) 'Of Digital Scholarly Editions and Building Tools.' Paper presented at *DiXiT 3 / ESTS 2016. Digital Scholarly Editing: Theory, Practice, Methods.* 5-7 October 2016: University of Antwerp, Belgium.

Apollon, D., Bélisle, C., and Régnier, P. (2014) *Digital Critical Editions.* Urbana, Chicago, and Springfield: University of Illinois Press.

Robinson, P. (2016) 'Project-based digital humanities and social, digital, and scholarly editions' *Digital Scholarship in the Humanities* Advance Access published September 18, 2016

Ruecker, S., Radzikowska, M., and Sinclair, S. (2011). *Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing.* Surrey: Ashgate Publishing, Ltd.

Sahle, P. (2014) 'Criteria for Editing Scholarly Digital Editions, version 1.1.' *I-D-E*. Available online on http://www.i-d-e.de/publikationen/weitereschriften/criteria-version-1-1/ (accessed on 1 November 2016; first published September 2012; last updated June 2014).

Sahle, P.(2016) 'A Catalog of Digital Scholarly Editions'. *Digitale Edition.* Available online on http://www.digitale-edition.de/ (accesed on 1 November 2016; last updated 19 May 2016).

W3C (2016) 'Accessibility, Usability, and Inclusion: Related Aspects of a Web for All'. *W3C.* Available online on https://www.w3.org/WAI/intro/usable (accessed on 31 October 2016; first published March 2010; last updated 6 May 2016).

Building Capacity for Digital Scholarship & Publishing: Three Approaches from Mellon's 2014-2015 Scholarly Communications Initiative

Sara Sikes
sara.sikes@uconn.edu
University of Connecticut, United States of America

Maria Bonn
mbonn@illinois.edu
University of Illinois at Urbana-Champaign
United States of America

Elli Mylonas
elli_mylonas@brown.edu
Brown University, United States of America

The Andrew W. Mellon Foundation's 2014-2015 Scholarly Communications Initiative funded more than thirteen projects of various sizes and orientations as part of an effort to strengthen the scholarly monograph publishing ecosystem in a time of increasing disruption. While the projects are seemingly divergent in their approaches, a recent report from Simon Fraser University, "Reassembling Scholarly Communications: An Evaluation of the Andrew W. Mellon Foundation's Monograph Initiative" (May 2016) identifies points of thematic alignment and overlap. Many of the funded projects are explicitly based in university presses, with the goal of either enhancing existing monograph programs or developing digital capacity where little or none exists. But three projects, located at University of Illinois, Brown University, and University of Connecticut, are instead focused on exploring new forms for scholarly expression and developing models for faculty digital publication outside the traditional press framework.

The "Building Capacity for Digital Scholarship & Publishing" panel brings together representatives from each of these three projects to investigate unique and complementary dimensions of their work. Presenters include Maria Bonn from the University of Illinois, Elli Mylonas from Brown University, and Sara Sikes from the University of Connecticut. The session will address the development of project outcomes and encourage participation in a discussion about large-scale shifts in structural and cultural approaches to faculty-led digital scholarship production and publication. Rather than building new publishing platforms, these three Mellon-funded projects are focused on exploring workflows and work cultures suited to the creation of multimodal scholarly communications. The University of Illinois will explore how current tools align with scholars' publishing requirements and address the gap between scholars' needs and the existing publishing systems for digital scholarship. At both Brown University and University of Connecticut, the scholarly publications initiatives are focused on developing an infrastructure to support digital scholarship projects, facilitating new workflows and work cultures, and revising criteria for tenure and promotion.

Publishing Without Walls

Maria Bonn

"Publishing Without Walls" is a Mellon-funded initiative at the University of Illinois, led by the University Library in partnership with the School of Information Sciences, the department of African American Studies, and the Illinois Program for Research in the Humanities. The objective is to develop a model for library-based publishing services that can be adopted broadly by other academic libraries to address scholars' emerging needs in a contemporary publishing environment. Both an embedded research effort and programmatic development are strategically designed to address known gaps within the current landscape: the gap between what and how scholars want to publish and what existing systems of print publishing can afford; the gap between the everyday practices of humanities scholars and the high-level tools that exist for digital scholarship; and the gap between digital scholarship and publishing at resource-rich institutions as opposed to at Historically Black Colleges and Universities.

Most experiments with new forms of and models for digital publishing focus on parts of the process rather than taking a holistic, scholar-centric view of publication: for example, tools are often developed with only an abstract sense of scholarly requirements, and new funding models are suggested without exploring whether and how existing systems meet scholars' needs. In order to guide the development of a more holistic and widely shareable service model, the University of Illinois is undertaking a significant qualitative study on how emerging services and tools align with scholars' publishing needs and using the findings to shape and inform the development of tools and services for Publishing Without Walls.

The Role of the University in Scholarly Communication

Elli Mylonas

The purpose of Brown University's Mellon-funded digital scholarship initiative is to establish an infrastructure to support the development and publication of digital scholarly monographs. Anchored in the University Library and the Office of the Dean of the Faculty, this digital publishing initiative extends the University's mission of supporting and promoting the scholarship of its faculty, while also playing a role in shaping the future of digital scholarship in the humanities more broadly.

Brown will first address questions about the evaluation of digital scholarship within the academic validation process. The initiative promotes the creation of language and criteria at the department level upon which digital scholarship will be evaluated for tenure and promotion purposes, and this process is, as of Spring 2017, substantially complete. With these guidelines established the initiative is now supporting two digital scholarly publications. Mellon grant funds have been used to augment the library's staff, adding new positions including the Digital Scholarship Editor, a Designer for Online Publications, and graduate- and undergraduate-level researchers. Funds from the grant are also designated to bring in speakers who are engaged in digital projects or other efforts to introduce digital scholarship and publication into the formal academic process in order to spark discussions around best practices and innovative approaches. By tackling first the issue of evaluation and then assisting with the creation of innovative, digital-first scholarship as well as innovative digital publication, the initiative will remove barriers in place at Brown and potentially serve as an example to other institutions that are engaging with emerging forms of writing, publication, and dissemination.

The Scholarly Communications Design Studio

Sara Sikes

The Digital Media & Design Department at the University of Connecticut (UConn), in collaboration with the University Library and UConn Humanities Institute, is implementing Greenhouse Studios | Scholarly Communications Design at UConn. Greenhouse Studios will facilitate a design-based, inquiry-driven, collaboration-first workflow that addresses the divided processes and counter-productive labor arrangements that have complicated scholarly communications in the digital age. Even as the scholarly communications field pursues the opportunities presented by digital technology, its routine operations remain anchored in print-centric regimens. For those striving to evolve digital scholarship production and publication in the Internet age, particularly as it bears upon long-form scholarship, there is compelling need to productively disrupt and reconfigure the processes and work cultures that have naturalized around the production of printed products.

The proposed approach of Greenhouse Studios is undergirded by design-thinking methodologies that foster divergent creativity in generating and developing ideas directed toward solving complex problems while keeping end recipients of a project squarely in mind. Design thinking practices do not, in and of themselves, address cultural or workflow hierarchies, therefore a refined design process model has been established so that work begins with a prompt, an inquiry-in-common, that is set before a project team. In the design process model for the Mellon initiative undertaken at UConn, the potential audiences of a publication are explored at the very outset of a project, in specific rather than vague terms, as an essential element of the workflow. The team members hail from the areas once conceived of as the transactional links in the scholarly communications process, including librarians, web developers, editors, and designers, and they are assembled as a collaborative unit at the very outset of each undertaking. The power of the values and ways of working to be instantiated through the Greenhouses Studios places continuous, equitable communication between all kinds of scholarly communications labor at the heart of its mission.

In its broadest sense, "scholarly communications" is conceived of as communication of scholarship to specific audiences in a manner designed to make the expression of knowledge informative, useful, and relevant to that particular community. By employing the use of multimodal communications, researchers and their scholarship are connected to broad peer or public constituencies, and a work may be expressed in multiple form, each tailored to a different type of audience. As we look beyond inward-focused text, the visual, aural, and interactive forms of expression open pathways into understanding complex material that print alone cannot.

Bibliography

Maxwell, J., Borodini, A. and Shamash, K. (May 2016) "Reassembling Scholarly Communications: An Evaluation of the Andrew W. Mellon Foundation's Monograph Initiative." Report. Canadian Institute for Studies in Publishing. Vancouver, BC.

Sustaining and Scaling the Digital Liberal Arts

Lee Skallerup Bessette
lee.bessette@gmail.com
University of Mary Washington
United States of America

Kristen Eshleman
kreshleman@davidson.edu
Davidson College, United States of America

Caitlin Christian Lamb
cachristianlamb@davidson.edu
Davidson College, United States of America

Siobhan Senier
siobhan.senier@unh.edu
University of New Hampshire, United States of America

Paul A. Youngman
youngmanp@wlu.edu
Washington & Lee University, United States of America

When we invoke the Digital Liberal Arts, we are not so much proposing a field that is by definition distinct from Digital Humanities—after all, both terms are notoriously subject to endless re-definition and contestation. Instead, we mean to strategically (and politically) call attention to the kinds of digital scholarship and pedagogy that are being conducted outside of traditional Research 1 institutions with well-funded DH centers. What conjoins our papers, then, is not so much that they present a series of new case studies in teaching with the digital: many people, we agree, could do that. Rather, we intend to prompt discussion and reflection about the kinds of infrastructures that are necessary (and perhaps not necessary) to produce sustainable and meaningful digital scholarship. Early on, Pannapacker proposed that small liberal arts institutions and programs might be uniquely positioned for rapid, more cost-effective innovation, because we have "shallower administrative hierarchies and less institutional inertia." (Pannapacker 2013) All of us hail from programs that "share a culture of faculty-student collaborative research, which translates perfectly into the project-building methods of the digital humanities" (Pannapacker 2013).

For these reasons, we are pondering the differences between large university DH and small college DH: a diffuse, decentralized approach to DH versus a more systematic and integrated one. Our varied papers share several connecting themes: collaboration across disciplines, roles and institutions; and the central place of pedagogy. The idea of pedagogy is critical to the mission of liberal arts colleges, and thus will feature prominently in this discussion. We place pedagogy at the center of all our work. There is much that larger institutions can learn from this discussion, including ways to make DH more student-centered and pedagogically oriented. Without romanticizing student-centered projects (we will also address some of the challenges for public scholarship with such a variety of skills at the table) we want to explore how involving students in the co-production of knowledge – digitally mediated and publicly presented – shapes and reshapes what is possible under "DH."

Bibliography

Pannapacker, W. (2013). "Stop Calling It 'Digital Humanities'". *The Chronicle of Higher Education.* 18 February. Web. http://www.chronicle.com/article/Stop-Calling-It-Digital/137325/.jobs__topjobs-slider

Mapping 20th Century America

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Jason Heppler
jason.heppler@gmail.com
University of Nebraska – Omaha
United States of America

Robert Nelson
rnelson2@richmond.edu
University of Richmond, United States of America

Over the last several years, spatial humanities have grown in prominence. Initiatives such as the University of Virginia's Institute for Enabling Geospatial Scholarship and Stanford's Spatial History Project have signaled and begun exploring the impact of the spatial turn. Writing for UVA's institute, historian Jo Guildi states, "The spatial turn represents the impulse to position these new tools [i.e. GIS] against old questions" (Guildi, 2010). In this panel, we challenge the idea that the spatial turn simply asks old questions and rather argue that it helps pose new as well as answer old questions in twentieth-century American history. In particular, we employ the idea of deep mapping as theorized by historian David Bodenhemer to discuss new scholarship and directions in digital spatial analysis. We also explore the role public humanities plays in framing these projects. The 45-minute panel entitled "Mapping the 20th Century United States" will focus on the role of the spatial analysis in the digital, public humanities and its impact on historical scholarship.

Robert Nelson will begin the panel with "Reckoning with Redlining: Public Engagement with 'Mapping Inequality.'" The Mapping Inequality project provides unparalleled access to the infamous redlining maps and area descriptions created by the Home Owners' Loan Corporation during the Great Depression. Accessed by tens of thousands of visitors in the first two weeks following its release, this paper will draw upon hundreds of contributions to a public conversation about redlining and urban inequality around "Mapping Inequality" in social media and the comment threads of press the project has received. Nelson will analyze the design decisions he and his colleagues made to prompt public engagement with these maps and the history of redlining. He will also critically use the reaction to the project to critically assess how successful this public-facing digital humanities project has been in prompting productive conversations about redlining and urban and racial inequalities.

Continuing with our exploration of the 1930s, we will turn to Taylor Arnold and Lauren Tilton who will discuss combining archives spatially in order to produce new knowledge about and public access to documentary expression in the era. They will focus on layering the Federal Writer's Project, which documented through text the life histories of thousands of Americans, with 170,000 photographs from the Farm Security Administration-Office of War Information. Placing these archives for the first time in conversation, their deep maps incite new questions about the role of the federal government in documenting the lived experiences of Americans during the Great Depression and the types of representation produced.

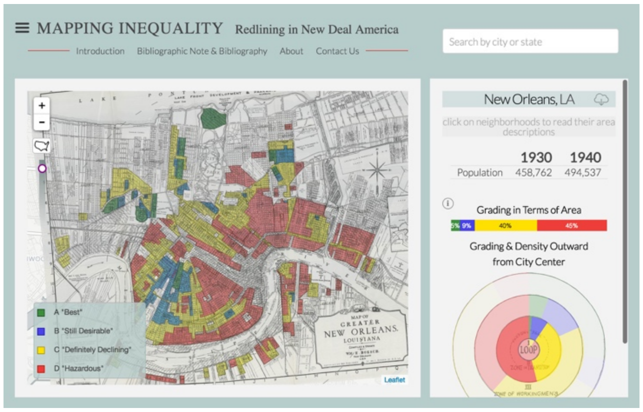
Next we will turn to the work of Jason Heppler on post-war America. Silicon Valley represented one of the twentieth century's greatest modernizations of urban space. Beginning in the 1950s, the formation of a new high tech suburbanism led the Valley to become identified not only with a center of hopeful possibility as the Industrial Age industries of the Midwest and Northeast began to decline, but also gave expression to an environmental politic that attempted to reconcile an environmentally conscious pursuit of the American Dream. Yet the claim for high tech's "clean" industrialization fell short as environmental concerns -- ranging from controlling growth to widespread chemical contamination of water supplies -- reshaped discussions about public and private space. Deep maps help explore the transformation of urban space over time.

Along with addressing the role of spatial analysis in cutting edge humanities scholarship, each paper will outline which technologies they are using along with their possibilities and challenges. In particular, Robert Nelson will address a cutting edge spatial toolkit the University of Richmond Digital Scholarship Lab and Statmen Design are developing for use across the digital humanities. The panelists will also discuss the role of collaboration, the process of developing cross-institutional partnerships and designing for public audiences.

Reckoning with Redlining: Public Engagement with "Mapping Inequality"

Robert Nelson

In keeping with the conference theme of access and its emphasis upon public-facing scholarship, this presentation will reflect upon hundreds of comments and several conversations from the lay public about "Mapping Inequality". A collaboration of teams at four universities, "Mapping Inequality" currently includes nearly all of the more than 150 "security maps" and nearly 10,000 "area descriptions" created by the Home Owners' Loan Corporation during the Great Depression. These maps assessed mortgage risk for thousands of neighborhoods in U.S. cities large and small on a scale of "A" to "D". "A" neighborhoods were deemed "best," presenting minimal risks for banks and lenders; "D" neighborhoods were deemed "hazardous" for mortgage financing.



New Orleans, LA

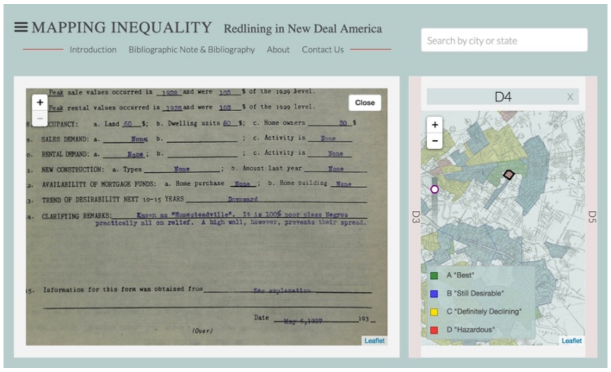
click on neighborhoods to read their area descriptions

	1930	1940
Population	458,762	494,537

Grading in Terms of Area: 5% 9% 40% 45%

Grading & Density Outward from City Center

These grades were explicitly racialist and racist. HOLC's survey instruments asked local agents to quantify the "infiltration of" undesirable populations of African Americans and immigrants. To cite just a few examples, a small subsection of a Tacoma neighborhood was graded "D" though otherwise identical to the surrounding "B" neighborhood because "Three highly respected Negro families own homes and live in the middle block of this area facing Verde Street. While very much above the average of their race, it is quite generally recognized by Realtors that their presence seriously detracts from the desirability of their immediate neighborhood." Proximity to black neighborhoods was enough to impact HOLC's risk assessment. A subsection of a neighborhood in Richmond was graded "C" rather than "B" because "Respectable people but homes are too near negro area D2." In contrast, a Camden neighborhood kept an"A" grade despite bordering an African American neighborhood, but only because "High walls separates this section from the colored area to the south" that effectively prevented their "spread."



Peak sale values occurred in 1928 and were 103 $ of the 1929 level.
Peak rental values occurred in 1928 and were 103 $ of the 1929 level.
8. OCCUPANCY: a. Land 60 $; b. Dwelling units 60 $; c. Home owners 30 $
9. SALES DEMAND: a. None; b. ; c. Activity is None
10. RENTAL DEMAND: a. None; b. ; c. Activity is None
11. NEW CONSTRUCTION: a. Types None; b. Amount last year None
12. AVAILABILITY OF MORTGAGE FUNDS: a. Home purchase None; b. Home building None
13. TREND OF DESIRABILITY NEXT 10-15 YEARS Downward
14. CLARIFYING REMARKS: Known as "Homesteadville". It is 100% poor class Negros practically all on relief. A high wall, however, prevents their spread.

15. Information for this form was obtained from See explanation

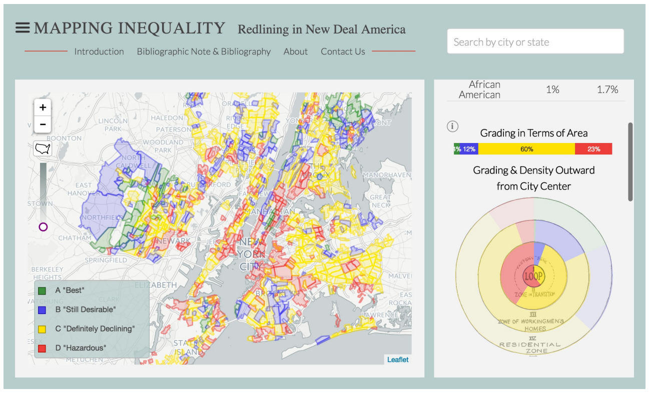
Date May 6,1937 193

(Over)

These grades had real consequences. Through this HOLC program, the federal government reinforced redlining as a best practice within the real estate industry, in effect cutting off hundreds of thousands of African Americans off from equitable access to mortgage financing and thus homeownership, which arguably was the most significant mechanism of familial wealth accumulation in twentieth-century America. While of course it is by no means the only or even primary cause, this redlining program helped to contribute to generational wealth disparities between white and black Americans, where today the median wealth of white households is a shocking 13 times that of black households.

We designed "Mapping Inequality" not only with researchers but activists and the general public in mind. More than 150 of the HOLC maps have been georectified (nearly all of them, though we still have a few to add and undoubtedly a few more will surface), and polygons for each neighborhood added. The site is location aware and asks new users if they want to view their, or alternately the nearest, city. The opacity of the raster maps can be adjusted to help viewers connect the grades to the contemporary cityscape.

Nearly all of the neighborhoods polygons can be clicked to read the area description survey. In short, we designed it hoping to encourage viewers to grapple with the materials related to their own localities and to prompt them to make connection to the present.



The introduction and other contextual materials on the site convey the authors' collective assessment that "New Deal era housing policies ... helped set the course for contemporary America." We also include a visualization inspired by Ernest Burgess's concentric circle theory to suggest that HOLC policies definition of the interior of cities as "slums" functioned as a self-fulfilling prophecy. Nevertheless, the site prioritizes access, exploration, and reuse of these important primary source materials. We do not provide these materials completely without commentary, but through our design choices we do facilitate and encourage relatively direct engagement with HOLC's maps and surveys.

All of these materials were available in the National Archives. While materials for many cities had been digitized, to date there has been no comprehensive collection let along one offering the functionality of "Mapping Inequality." We have no doubt that these materials will be useful to researchers--not just historians but economists, urban planners, artists, medical doctors, etc.--and will facilitate a far more nuance understanding of HOLC and its consequences. We also have abundant evidence that these materials are a boon for activists working on fair housing and other social justice causes as well.

While we're excited about this, we intentionally developed the site as a public history project that aimed to spark conversations about wealth and racial inequality in American cities past and present. By that measure, the project so far has been a success. Two and a half weeks after being released, the map has received about 44,000 visits and been the subject of online coverage from NPR, National Geographic, Slate, CityLab, FastCo., Forbes, and Curbed, all of which have narrated the state's role in fostering redlining and wealth inequality.

In the comment section of these stories, in stories in local news sources, and in social media there has been a broad-ranging conversation about wealth and racial inequalities. On one of the spectrum, some respondents have been dismissive of the project, suggesting that this happened 80 years ago and is a remnant of the past that has little relevance today; one person notably characterizing the site as nothing more than "historical racism porn." Others have responded to such comments that this is important inasmuch as many of these 80-year old maps resemble the racial and class landscapes of America today and that the government's role in reinforcing redlining and racial disparities of wealth isn't widely understood.

Beyond these arguments about the impact of HOLC and relevance of redlining for understanding inequality in twenty-first-century cities, some of the most interesting and revealing comments have come from people for whom the maps have prompted reflection upon their own family histories. "I grew up in Detroit in the late 50's and 60's," one man wrote. "My address was 20400 Monte Vista, the corner Monte vista and Norfolk. Two streets east, starting at the corner of Birwood and 8 mile, was a wall. The wall was 12-15 feet high, made of grey concrete blocks and ran behind the homes towards 7 mile, extending to an abandoned army base at the corner of Pembrook and Birwood. The wall was built to divide the neighborhoods, one side was all African Americans, The other all Caucasian. My mother lived on one side of the wall, it was all African American. She once (just once) told me that one could hear the White families on the other side of the wall talking, see them occasionally if a ball came over the fence and they asked for it, most times they did not. One had no contact, ever. The wall is still there, physically and emotionally." While so far this story is rather exceptional in its detail, it has prompted us to think about the possibility of using the site to solicit and collect stories about the consequences of redlining and segregation on particular individuals, families, and communities.

Given that "Mapping Inequality" has been at the center of several online conversations and hundreds of comments on websites two and a half weeks after it was first released, I'm optimistic that it will continue to occasion more conversations about the role of racism, redlining, and the state in inequalities of wealth in American cities. This presentation will provide an opportunity to critically reflect upon these materials and gauge the success and failures of this particular digital humanities project and perhaps the digital humanities more generally in informing socially and politically important public conversations. This presentation will also reflect upon the pros and cons of interpretive framing in digital humanities projects aimed at the public.

Mapping the Federal Writers Project

Lauren Tilton and Taylor Arnold

“Mapping the Federal Writers Project” will explore the role and implementation of deep mapping and spatial analysis in interpreting and understanding documentary expression in 1930s America. As Bodenhamer argues, deep maps are “visual, time-based genuinely multimedia and multilayered” (Bodenhemer, 11). These maps allow for nuanced spatial analysis in the service of new humanities questions and arguments. We will focus on the application of these concepts in a new extension of Photogrammar (photogrammar.yale.edu), a digital and public humanities project focused on print and visual culture in 1930s America.

Photogrammar (photogrammar.yale.edu) uses methods from the digital humanities and digital resources to further contextualize and open new avenues of research into the federal project and documentary record of the era. In its current version, Photogrammar maps 170,000 photos from the Great Depression and World War II that comprise the United States Farm Security Administration and Office of War Information (FSA-OWI) photographic archive. Importantly, the collection includes some of the most prominent documentary photographers of the 20th century including Dorothea Lange and Walker Evans. Users can explore the collection through interactive maps through the use of spatial analysis; search by photo captions, photographer and time through text analysis; and browse by color and the faces depicted in the photographs through the use of image analysis. This new stage - funded by the American Council of Learned Societies- involves adding a new layer to Photogrammar - the Federal Writer’s Project (FWP), which funded writers to capture and describe the complexities of American life during the Great Depression (Couch, Hirsch, Mangione, Penkower, Stewart). Dozens of writers were sent to document through words the impact of the great depression on people's lives across the country. Prominent literary scholars such as Nelson Algren, known for The Man with the Golden Arm , and Ralph Ellison, known for Invisible Man. In the process, they asked people to provide their life histories pioneering the practice of oral history, a critical methodology in the field of history and in the humanities more broadly.

Using deep mapping to expand our understanding of 1930s America, Photogrammar is creating links across archives in order to place the FSA-OWI in the larger federal effort to document America during the Great Depression. Merging collections from the University of North Carolina - Chapel Hill Libraries and the Library of Congress, the FWP includes over 4,000 life histories. Interviews are being plotted on a new geographical layer allowing search by space and time. For example, a user will be able to follow an interviewer as they move across a state and the country to collect oral histories in the same way they can now follow documentary photographs like Dorothea Lange and Walker Evans. Users will be able to search the new FWP layer independently or along with the geographical layer of FSA-OWI photographs and photographers. As a result, they will be able to compare the oral histories to the photographs taken in the same area allowing user to compare and contrast the documentary record created and funded by the federal government. As well, the new and cleaned transcripts created over this year are allowing for refined search functionality including faceted browsing and full text search. We are also experimenting with the role Natural Language Processing techniques such as Named-Entity Recognition to create new ways to browse the collection (Finkel and Manning 2005). In all, users will be able to explore the FWP and FSA-OWI spatially, temporally and through faceted searching allowing the public to explore the broader documentary record of the era relationally through deep mapping.

The second half of the paper will focus on methodology. We will start by discussing the potential benefits and potential difficulties of cross-institution collaboration in the cleaning and processing of data. In our experience, working with institutionally and spatially separated groups requires careful planning, but this extra up-front work improves both the general workflow and final products. We will touch on what scholarly and which technical questions guided our creation of a database schema for inputting metadata from the Federal Writer's Project. We take into consideration theoretical work regarding the creation of "smart data", best practices regarding TEI markup (Schöch, TEI Initiative on Libraries), and principles for creating normalized database tables (Codd 1971). The discussion will culminate in showing how these carefully curated data sources are made interactive and public on our website. The geographic data are plotted using custom layers created in CARTO (formerly CartoDB), giving a great deal of interactivity out of the box. Specifically, we will extrapolate on how we designed the interactivity to enhance the other data collections on the Photogrammar website, to make new arguments and pose new questions about documentary expression in 1930s America, to realize the principles of deep mapping and to engage with various publics.

Mapping Silicon Valley

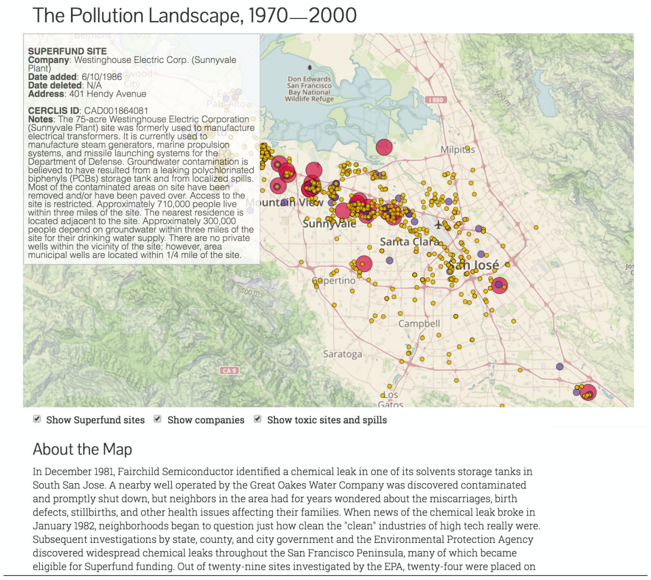
Jason Heppler

“Mapping Silicon Valley” explores the role of spatial history in the urban environment of post-World War II Santa Clara Valley. Silicon Valley is the product of competing landscapes. The geographer D. W. Meinig refers to landscapes as “a naïve acceptance of the intricate intermingling of physical, biological, and cultural features which any glance around us displays" (Meinig, 1979). Wildlife refuges, fenced military installations, city and neighborhood districts, and polluted sites all hold definitions on the land. Historian Richard White has referred to this as "hybrid landscapes," where cultural ideologies clash over conflicting uses of natural resources. The hybrid landscape is neither purely wild nor purely built, but instead a construction of natural and cultural systems that shape and create place (White, 2004). People define places by embedding ideas on the landscape. In cities, urban planners lay down grids of roads, zones, and regulations that divide cities along labor, leisure, and consumption, thus imbuing certain places with particular meaning. Landscapes, as Meinig notes, are “a great exhibit of consequences," and are “symbolic, as expressions of cultural values, social behavior, and individual actions worked upon particular localities over a span of time" (Meinig, 1979).

By viewing Silicon Valley through the lens of landscapes and space, I argue for the importance of place in shaping a suburban vision of what urban historian Margaret O'Mara has called high-tech urbanism. Silicon Valley has come to represent the future of post-industrial economic development. Places as varied as Atlanta, Georgia; Philadelphia, Pennsylvania; Cleveland, Ohio; Omaha, Nebraska; Bangalore, India; Mission Hills in the Guandong Province of China; and Shenzhen, China, have looked to Silicon Valley as a model for economic and urban revitalization through high-tech economic development. Indeed, high tech is often drenched in green--from high-tech office campuses to "smart cities" that promise to transform work, leisure, transportation, and urban space into a more sustainable future.

The work of this high-tech landscape has been decades in the making and has come with high environmental costs, despite the promise of clean and green cities. Silicon Valley epitomized the trend of conflating a lack of smokestacks as a proxy for sustainable industrial development. High-tech landscapes centered around industrial research and scientific industry promised growth without pollution, but that promise was an impossible standard.

"Mapping Silicon Valley" is a broad discussion of three map-centric projects that have moved through different stages. The first set of maps were data driven thematic maps, produced largely during the course of my dissertation research. These maps were created largely out of a desire to understand the transforming landscape in Silicon Valley, from city growth and conflicts over urban space to the widespread presence of pollution and neighborhoods most threatened by toxic chemicals. The first section of this paper will reflect on the methodological underpinnings of these maps and their application to environmental humanities, while also discussing some of the potential shortcomings and enhancements that would make these maps more useful for historical research.



The Pollution Landscape, 1970—2000

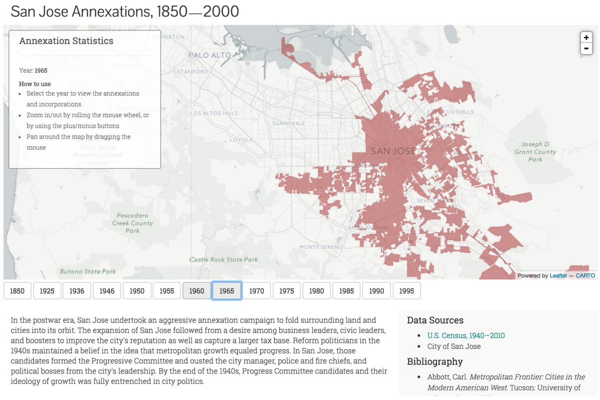
SUPERFUND SITE
Company: Westinghouse Electric Corp. (Sunnyvale Plant)
Date added: 6/10/1986
Date deleted: N/A
Address: 401 Hendy Avenue

CERCLIS ID: CAD001864081
Notes: The 75-acre Westinghouse Electric Corporation (Sunnyvale Plant) site was formerly used to manufacture electrical transformers. It is currently used to manufacture steam generators, marine propulsion systems, and missile launching systems for the Department of Defense. Groundwater contamination is believed to have resulted from a leaking polychlorinated biphenyls (PCBs) storage tank and from localized spills. Most of the contaminated areas on site have been removed and/or have been paved over. Access to the site is restricted. Approximately 710,000 people live within three miles of the site. The nearest residence is located adjacent to the site. Approximately 300,000 people depend on groundwater within three miles of the site for their drinking water supply. There are no private wells within the vicinity of the site; however, area municipal wells are located within 1/4 mile of the site.

☑ Show Superfund sites ☑ Show companies ☑ Show toxic sites and spills

About the Map

In December 1981, Fairchild Semiconductor identified a chemical leak in one of its solvents storage tanks in South San Jose. A nearby well operated by the Great Oakes Water Company was discovered contaminated and promptly shut down, but neighbors in the area had for years wondered about the miscarriages, birth defects, stillbirths, and other health issues affecting their families. When news of the chemical leak broke in January 1982, neighborhoods began to question just how clean the "clean" industries of high tech really were. Subsequent investigations by state, county, and city government and the Environmental Protection Agency discovered widespread chemical leaks throughout the San Francisco Peninsula, many of which became eligible for Superfund funding. Out of twenty-nine sites investigated by the EPA, twenty-four were placed on



San Jose Annexations, 1850—2000

Annexation Statistics

Year: **1965**

How to use

- Select the year to view the annexations and incorporations
- Zoom in/out by rolling the mouse wheel, or by using the plus/minus buttons
- Pan around the map by dragging the mouse

In the postwar era, San Jose undertook an aggressive annexation campaign to fold surrounding land and cities into its orbit. The expansion of San Jose followed from a desire among business leaders, civic leaders, and boosters to improve the city's reputation as well as capture a larger tax base. Reform politicians in the 1940s maintained a belief in the idea that metropolitan growth equaled progress. In San Jose, those candidates formed the Progressive Committee and ousted the city manager, police and fire chiefs, and political bosses from the city's leadership. By the end of the 1940s, Progress Committee candidates and their ideology of growth was fully entrenched in city politics.

Data Sources

- U.S. Census, 1940—2010
- City of San Jose

Bibliography

- Abbott, Carl. *Metropolitan Frontier: Cities in the Modern American West.* Tucson: University of

The second mapping project is oriented around digital public history. Called Silicon Valley Historical and built on the Curatescape platform, the project seeks to collect archival material and narrate the importance of specific places to the Valley's history. Contributions to Silicon Valley Historical are not solely driven by scholarly contributions, but also rely on contributions by students and volunteers in close association with area universities, colleges, historical associations, and historical societies. The material contained in Silicon Valley Historical is meant, in part, to step away from the business-centric stories so often associated with Silicon Valley and consider more fully the urban spaces that were affected by the growth of this high-tech region. Still in its early stages of planning, this section will discuss the challenge of working with community partners and developing a sustainable and scalable digital history project that seeks to serve both the community it studies as well as students and scholars who will find the project useful.

The final mapping project, still under planning and a partnership with the Stanford Spatial History Project, is tentatively titled From Orchards to Suburbs: Changing Landscapes in Silicon Valley and will represent the most technological and research heavy aspects of the project. As I investigate the politics surrounding the creation of place, this project will allow for the spatial exploration of zoning laws, general plans, government reports, and city council meeting minutes. The current design envisages the ability to navigate through a map and, depending on the viewport, presenting a list of primary sources available for reading about particular places in Silicon Valley. These will be accompanied by some computational and statistical tools for doing text analysis to uncover more about the kinds of conversations happening about particular places in the city and how they are being thought about.

Collectively, "Mapping Silicon Valley" will critically reflect on these projects and their evolution as research and public history projects. The paper will further delve into the opportunities for deep mapping and interactivity for exploring the changing landscapes of Silicon Valley.

Bibliography

Bodenhamer, D. (2013). *History and GIS: Epistomologies, Considerations and Reflections.* Springer.

Codd, E. F. (1971) "Further Normalization of the Data Base Relational Model". (Presented at Courant Computer Science Symposia Series 6, "Data Base Systems", New York City, May 24–25, 1971.) IBM Research Report RJ909 (August 31, 1971).

Couch, W. T., ed. (1939) These Are Our Lives. (University of North Carolina Press, 1939).

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370

Guildi, J. (2010) “What is the Spatial Turn?”.Available online: http://spatial.scholarslab.org/spatial-turn/

Hirsch, J. (2003) Portrait of America: A Cultural History of the Federal Writers’ Project: A Cultural History of the Federal Writers’ Project. (University of North Carolina Press, 2003).

Mangione, J. (1996) *The Dream and the Deal.* (Syracuse University Press).

Meinig, D., ed., (1979). The Interpretation of Ordinary Landscapes (New York: Oxford University Press), 2

Penkower, M. N. (1977). *The Federal Writers’ Project: A Study in Government Patronage of the Arts.* (University of Illinois Press).

Schöch, C. (2013). “Big? Smart? Clean? Messy? Data in the Humanities.” *Journal of Digital Humanities* 2, no. 3.

Stewart, C. A. (2016). *Long Past Slavery: Representing Race in the Federal Writers’ Project.* (University of North Carolina Press, 2016).

TEI Initiative on Libraries (2011). “Best Practices for TEI in Libraries.” *Text Encoding Initiative*. October 2011. Available online at: http://www.tei-c.org/SIG/Libraries/teiinlibraries/

White, R. (2004). “From Wilderness to Hybrid Landscapes: The Cultural Turn in Environmental History,”The Historian 66 (September 2004): 562–664.

Schrifttanz Zwei

Susan Wiesner
swiesner@umd.edu
University of Maryland, United States of America

Rommie L. Stalnaker
rstalnaker81@gmail.com
Independent Scholar

Stephen Ramsay
sramsay.unl@gmail.com
University of Nebraska-Lincoln, United States of America

Brian Pytlik Zillig
bzillig1@unl.edu
University of Nebraska-Lincoln, United States of America

This proposed panel will use a phenomenological perspective as well as a production-oriented approach to the work/research on which we are collaborating (‘Schrifttanz zwei’). Although digital technologies support this multi-disciplinary project that combines archival research, dance choreography, music composition, animation creation, and video projection with a goal of performance production, this panel will present our collaborative process and how we access and move through the digital and analog spaces in which we each work. The hope is that through this discussion (and performance event) attendees (and we as collaborators) will come to better understand the place of the Arts (Humanities writ large) in the Digital world.

The idea for Schrifftanz zwei began at the DH2016 conference where long-time DHers Steve Ramsay and Brian Pytlik Zillig presented their work with music and animation: ‘Picture to Score: Driving Vector Animations with Music in the XML Ecosystem’. Well known for their work within the Digital Humanities arena, their presentation demonstrated that the two had crossed boundaries into the nonverbal world of the arts. Two attendees who listened to their talk quickly became interested in collaborating. But these two, Rommie Stalnaker and research colleague Susan Wiesner, hail from the other side of the fence: they are Performing Artists who choreograph, perform, and conduct DH research using Dance. Sharing their desire to work with Ramsay and Pytlik Zillig, the four became inspired by the possibilities and began what has become an exciting collaboration challenged by three time zones, 3000 miles, and four personal processes for re-imagining a dance score created in 1927 by Irmgard Bartenieff, founder of the Laban/Bartenieff Institute for Movement Studies.

The original document from which we re-imagine the dance was discovered during the processing of an archival collection at the University of Maryland Special Collections in the Performing Arts. Bartenieff, a dancer and student of Rudolf Laban, brought Laban's theories about movement and his notation system to the USA in 1936 when she emigrated from Germany. Bartenieff began composing dances while studying with Laban in 1926-1929, yet she wasn't yet an expert in Schrifttanz, Laban's initial label for his notation system (Kinetography Laban, or Labanotation). Instead, she used her own hybrid system consisting of symbols, colour, and text to describe the dance movement and space. Better known for her work with Somatics, Laban's Effort/Shape theories, Motif writing, and Labanotation, the discovery of 4 choreographic works within Bartenieff's notebooks dated 1927-28 generated a desire to see them off the page (excerpt included here, Figure 1).

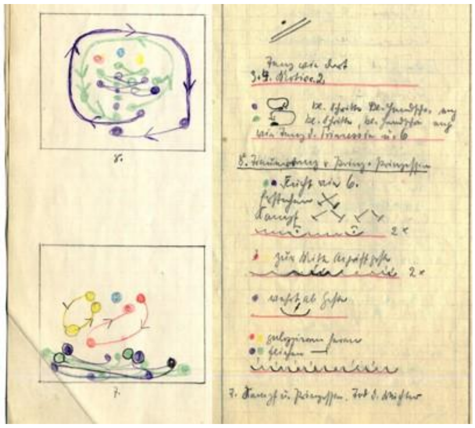


Figure 1: Excerpt from one of Bartenieff's notebooks

Because there are few particulars surrounding the work (who, what, when, where), we recognized that it is impossible to recreate it as it was without additional sources of information. Thus we decided to proceed with a re-imagining of the score to challenge our disciplinary approaches while giving voice to our individual creativity. For example, although the score is not written using Motif writing, nor Labanotation as we know them today, Bartenieff did call upon her knowledge of Laban Movement Analysis (LMA) by including a few elements in her textual descriptions. As the choreographers and dancers are all either CMAs or students of LMA we were able to understand those concepts integral to LMA. Also, we were able to support the movement choices made during the re-imagining by referring to Rudolf Laban's theory (from his 1926 *Choregraphie*), Bartenieff's personal movement style (referred to in archival papers as "Light"), and an analysis of the movement vocabularies used by German choreographers working in the 1920s.

Another factor in this project is the acknowledgement that researchers who attempt to use Laban theory and Notation in Computer Recognition and HCI continue to encounter what are often perceived as insurmountable issues. With the inclusion of LMA into Schrifttanz zwei (analyzing the performative product) we will not only use Bartenieff's work to inspire, but also to move closer to solving some of the issues faced by researchers using movement-based data. Thus, previous work on the ARTeFACT project (Wiesner and Stalnaker), which has long strived toward the use of Motif writing and LMA to enable automated tagging and retrieval of movement-based data, supports Schrifttanz zwei as yet another piece of the puzzle toward these goals.

So, too, does Schrifttanz zwei enable all team members to acknowledge our creative practices within a research framework. We each must interpret the verbal instructions and visualizations on some level through sound, visuals, and physical movement, i.e. the choreographic process must find a mutuality with Ramsay's musical composition and Brian's SVG videos inspired by Bartenieff's choreographic notes, graphics, and descriptions. Ramsay and Pytlik Zillig's recent artistic work uses Indigo, a program developed by Brian L. Pytlik Zillig for performing command-line stop-motion animation using Scalable Vector Graphics (SVG). Indigo produces thirty SVG files for each second of film, rasterizes them into JPEG images, and assembles them into H.264 high-definition video. Indigo animations can be quite simple or very complex. They may include text, shapes, colors, paths, layers, masks, and patterns. Indigo may be used to construct elaborate jointed characters that walk, talk, dance, or fly. Animations are modeled in XSLT and SVG, and can be programmatically synchronized with pre-existing audio using MusicXML metadata.

Schrifttanz zwei is admittedly an interdisciplinary artistic collaboration, but we would argue that the production of a work of art does not preclude the use of the digital; and indeed, Schrifttanz zwei includes born-digital elements (music and animation) intertwined with the born-human components and written/archived texts. Also, this collaboration is possible because of the prior work of the collaborators as it reflects the early phases of ARTeFACT and Ramsay and Pytlik Zillig's work with animation produced from digitized musical scores (Indigo). The proposed panel will address our collective and individual experiences in our art forms, as well as experiences using text and movement-based approaches in our DH research. Further this project is intended to create a Whole, where all voices and art forms share equal value with the supporting technologies, without privileging any one element. To accomplish this, we must negotiate within Digital Humanities AND the Arts. In fact, through this collaboration we have been made even more aware of the conversations surrounding definitions of the Digital Humanities, a topic we keep returning to during our collaboration. To wit: what is the place of the Arts in the Digital Humanities and what is required of a project to be aligned with the Digital Humanities? As DH artists as well as producers and users of digital technologies (e.g. Indigo, ARTeFACT, IDMove, etc.), we hope this panel/performance will provoke discussion and perhaps inspire others to find ways to access other 'outlier' disciplines through collaborative activities. Finally, as this collaboration constantly reminds us: "As technology and machines consume more and more of life, perhaps theater [read: dance] can help us remember what it means to act like a human." (Moore, 2016)

As an added component -- although the non-verbal is not included as an official ADHO language -- we plan to consider it as a communication method by allowing attendees to access their embodied knowledge through a brief non-verbal experience during the panel, including a request that some questions be asked and answered non-verbally.

Bibliography

Coartney, J. and Wiesner, S. (2009), 'Performance as Digital Text: capturing signals and secret messages in the media rich experience' in *Literary and Linguistic Computing* Special Edition, 24:2, June 2009.

Irmgard Bartenieff Papers, Special Collections in the Performing Arts, University of Maryland Libraries.

Laban, R. (1926), *Choregraphie* Jena: Eugen Diederichs Verlag.

Moore, Tracey. (2016) "Why Theater Majors Are Vital in the Digital Age". The Chronicle of Higher Education. 3 April. Web. http://www.chronicle.com/article/Why-Theater-Majors-Are-Vital/235925?cid=cp79

Simpson, T., Wiesner, S., and Bennett, B. (2014). 'Dance Recognition System Using Lower Body Movement' in *Journal of Applied Biomechanics* 30:1, February 2014.

Wiesner, S., Bennett, B., and Stalnaker, R. (2011). 'ARTeFACT Movement Thesaurus', White Paper, NEH Office of Digital Humanities.

Wiesner, S. and Stalnaker, R. (2016). 'Representing Conflict through Dance: using quantitative methods to study choreographic time, stage space, and the body in motion,' in *With(out) Trace: inter-disciplinary investigations into time, space and the body*, Dwyer, S., R. Franks and R. Green (Eds). e-book, Inter-Disciplinary Press, Oxford: United Kingdom.

Wiesner, S., Stalnaker, R. and Austin, A. (2016). 'Training the Machine: Movement and Metaphor' in *Embodied Performance: Design, Process and Narrative,* Oxford: Inter-Disciplinary Press.

Copyright, Digital Humanities, and Global Geographies of Knowledge

Vika Zafrin
vzafrin@bu.edu
Boston University, United States of America

Isabel Galina Russell
igalina@unam.mx
Universidad Nacional Autónoma de México, Mexico

Alex Gil
colibri.alex@gmail.com
Columbia University, United States of America

Padmini Ray Murray
p.raymurray@gmail.com
Srishti Institute for Art, Design and Technology, India

Copyright is controversial and murky. By any name, it refers to rights connected to original works. In some countries, emphasis is on the rights of authors or distributors of a work; in others, public good is given prominence. In most places it's a combination of the two. Significant commercial and personal interests are involved, and the advent of the internet has redefined distribution; and so copyright laws are constantly changing. Although the notion of intellectual property exists in almost all countries its significance and actual implementation varies as does the degree to which the counter balancing rights of access to information are implemented and supported.

In digital humanities, copyright is implicated in multiple ways. As with all scholarship, book and article publishing is fraught with rights transfers embedded in publishing contracts, and subsequent publisher practices that often impede scholarly progress. Other copyright issues are unique to DH. Who holds copyright to what part of a collaborative, web-based project? What is the difference, and what are the intersections, between having, protecting, and exercising legal rights on one hand, and being given proper credit on the other?

We propose a DH2017 panel to discuss these issues, with an eye toward helping DH practitioners make better informed decisions regarding the new knowledge they create. More importantly, though, we aim to spur a community conversation about copyright as an area in which digital humanists have agency as knowledge producers, not just as consumers. The rapid worldwide expansion of digital humanities work demands that we begin to deal with the complex tangle of rights around digital humanities knowledge production before others do it for us.

This conversation has not yet happened at the DH conference or in the field at large in a robust way. Copyright is mentioned usually as an issue around using third-party materials, an obstacle or a limitation for a particular project (Neuman, Greent, Unsworth 1997; Evenson 1999; Lord 1999; Ben-Porat, Reich, Behrendt 2002). We were unable to locate conference abstracts that addressed copyright and intellectual property issues for the new knowledge produced. Workshop-type events on DH and copyright tend also to take a more applied approach, on how third-party copyright affects the project. Two related conversations are taking place, though. One is about labor in academe at large and in DH in particular (Keralis, Burgess and Hamming, Anderson et al, Flanders); the other is about access to knowledge as a social justice issue—what access to scholarship means in terms of power distribution, perpetuating systemic inequalities in academe and outside it (Risam et al, Chenier, Faull et al). We think that our panel will contribute to this conversation. Considering questions of labor from the perspective of rights to knowledge produced potentially clarifies, and makes more expansive, our collective notion of where labor exists in a digital humanities project. And, thinking of the knowledge we produce in terms of rights we possess to it highlights our agency as individual contributors. This enables us to consider whether our individual and collective practice around author rights promotes or impedes our work's overall contribution to society.

We envision this panel beginning with a 10-12 minute presentation from each panelist on the topics described below, followed by a discussion with the audience.

Isabel Galina Russell will speak on the long tradition of Open Access publishing in Latin America and discuss how this may have impacted the way in which DH resources are produced, disseminated and published in this region. She will draw on examples of DH projects in the region and analyse the copyright situation, as well as present the results of conversations with DH creators and their attitudes and experiences with copyright in relation to their work.

Alex Gil will address the intersection between shadow libraries, digital humanities, vendor databases and digital libraries. Departing from the research work of the Piracy Lab and the Group for Experimental Methods in the Humanities at Columbia University, Gil will argue that our burgeoning global, hybrid republic of letters is being shaped in specific ways by the relationship of specific sectors of society to intellectual property. In a sense, Gil argues, architectures of humanistic knowledge are being produced by these different sectors, where the end product may resemble each other, but in large part due to the work of intellectual property laws, the labor conditions and social impact are quite different. The talk will conclude with a series of proposals for future humanistic research in the burgeoning area of what Alan Liu calls "critical infrastructure studies."

Padmini Ray Murray will address how the global hegemonies of knowledge production, ownership and circulation are challenged by creation and consumption practices in India, as embodied by the landmark ruling of the Delhi High Court in 2016, which dismissed suits filed by three international publishers who alleged that the circulation of photocopied material was an infringement of copyright. This will contextualise Ray Murray's consequent discussion of how both academic publishing and conventions of copyright are largely colonial legacies, and how this case marks a significant moment in the decolonisation of the Indian university and intellectual life. Ray Murray will demonstrate how practices in Indian language publishing might exemplify alternatives to dominant regimes, as well as how digital spaces and practices can and are already fostering an emergent model of humanities scholarship different than that found in the global North.

Vika Zafrin will give a brief overview of the current state of copyright in the United States as it relates to academic work from both the consumer/reader/user and the creator/author points of view. She will touch on some struggles academic libraries face in the current scholarly publishing climate, and describe strategies some institutions have adopted to improve the situation, including open access policies and an access-oriented approach to stewardship of digital (or digitized) scholarly materials given to libraries for archival and preservation. Zafrin will compare the institution- and funding-agency-level approaches to open access generally taken in the U.S. to those seen in some European countries (and in Europe as a whole via the Europeana project), highlighting differences and comparing their effectiveness while taking account what is possible in the respective political and legislative climates. She will briefly discuss some representative collaborations between U.S. researchers and colleagues worldwide, and their treatment of rights issues. Finally, she will make some suggestions for resources DH practitioners can use to make decisions about rights claims and attribution in their projects.

Bibliography

Anderson, K., et al. (2016) "Student Labour and Training in Digital Humanities." Digital Humanities Quarterly 10(1). http://www.digitalhumanities.org/dhq/vol/10/1/000233/000233.html. Accessed on 28 October 2016.

Ben-Porat, Z., Reich, S., Behrendt, W. (2002) "Organizing Multimedia Inter-textual Knowledge: New Tasks, Challenges and Technologies". In New Directions in Humanities Computing ALLC/ACH, July 23-28 2002, University of Tuebingen, July 23-28 2002 .http://lists.village.virginia.edu/lists_archive/Humanist/v15/0266.html. Accessed on 28 October 2016.

Burgess, H. J. and Hamming, J. (2011). "New Media in the Academy: Labor and the Production of Knowledge in Scholarly Multimedia." Digital Humanities Quarterly 5(3). http://digitalhumanities.org/dhq/vol/5/3/000102/000102.html. Accessed on 28 October 2016.

Chenier, E. (2014). "Oral History and Open Access: Fulfilling the Promise of Democratizing Knowledge." NANO: New American Notes Online 5, July 2014. http://www.nanocrit.com/issues/5/oral-history-and-open-access-fulfilling-promise-democratizing-knowledge. Accessed on 28 October 2016.

Evenson, J. (1999). "Electronic Archives: Creating a New Bibliographic Code". In International Humanities Computing Conference ACH/ALLC, June 9-13 1999, University of Virginia, USA. http://www2.iath.virginia.edu/ach-allc.99/. Accessed on 28 October 2016.

Faull, K., Jakacki, D., O'Sullivan, J., Earhart, A., Kaufman, M. (2016). Access, Ownership, Protection: The Ethics of Digital Scholarship. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 66-68.

Flanders, J. (2012) "Time, Labor, and "Alternate Careers" in Digital Humanities Knowledge Work." In: Gold, Matthew K., and Lauren F. Klein, eds. Debates in the Digital Humanities.

2012. http://dhdebates.gc.cuny.edu/debates/text/26. Accessed on 28 October 2016

Keralis, S. (2016) "Milking the Deficit Internship." In: Kim, Dorothy and Jesse Stommel, eds. Disrupting the Digital Humanities. Forthcoming. online at http://www.disruptingdh.com/milking-the-deficit-internship/. Accessed on 28 October 2016.

Liu, A. (2016). "Drafts for Against the Cultural Singularity (Book in Progress)." The Hyperarchival Parallax. . Web. 2 May. 2016. http://liu.english.ucsb.edu/drafts-for-against-the-cultural-singularity/. Accessed on 28 October, 2016.

Lord, L. (1999). "Keeping Our Word: Preserving Information Across the Ages". In: International Humanities Computing Conference ACH/ALLC, June 9-13 1999, University of Virginia, USA. http://www2.iath.virginia.edu/ach-allc.99/. Accessed on 28 October 2016.

Neuman, M., Green, D., Unsworth, J. (1997) "ACH Special Session: ACH and NINCH". In Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing ACH-ALLC, June 3 - 7 1997, Queen's University, Canada. http://web.archive.org/web/20020812020911/http://www.cs.queensu.ca/achallc97/. Accessed on 28 October 2016.

Risam, R., et al. (n. d.) "Access." In: Social Justice and the Digital Humanities. http://criticaldh.roopikarisam.com/criticaldh/access/. Accessed on 28 October 2016.

Smart Data Approaches to Exploring Independent Datasets across Disciplines, Media, and Perspectives for Research in the Humanities

Marcia Lei Zeng
mzeng@kent.edu
Kent State University, United States of America

James Lee
james.jlee@uc.edu
University of Cincinnati, United States of America

Research in the humanities has embraced the data-driven environment where advanced digital technologies have created the possibility of novel and hybrid methodologies. While the world is amazed by the many "V"s of Big Data (volume, velocity, variety, variability, veracity), the "V"alue of such data relies on the ability to achieve big insights from such data at any scale, great or small (Kobielus 2016). Smart Data can be considered trusted, contextual, relevant, cognitive, predictive, and consumable. In humanities research, Smart Data emphasizes the organizing and integrating processes from unstructured data to structured and semi-structured data, to make the big data smarter (Kobielus 2016, Schöch 2013).

In this panel, the interdisciplinary research teams from two universities will share their research findings and products as well as their experiences of using smart data approaches. As is the case with many other projects, the teams have to face the inherent challenges of converting heritage materials that were not machine-processable into digital datasets. Furthermore, the teams (again, like many others) have used common applications of available digital technologies and tools, such as GIS mapping, text encoding, fact mining, database construction, and visualization. In addition, the projects have employed sophisticated computer logics, the Linked Data models, network theory, and temporal-spatial data analytics, among others. However, the unique value of these projects lies in the exploration of independent datasets across disciplines, media, and perspectives. Some of such datasets were built from unstructured data and media, while some other datasets have been existed isolated because the values are hidden in the silos. By drawing on many types of data simultaneously and interactively in an unprecedented manner, the research findings and established resources help reveal the unknown-unknowns and interpret significant values.

To begin the panel, Dr. Marcia Zeng (Professor of information science) and Dr. Hongshan Li (Professor of history) will report on the project "Digital Humanities Research with Smart Big Data — A Network Framework of Innovation History" using the case of the Liquid Crystal Institute (LCI) at Kent State University (KSU), the birthplace of liquid crystal displays. Through the work of its faculty and alumni, LCI has had a significant impact on the way the world sees things –on our smartphones, tablets, and computer screens (Bos, 2015). By nature, innovations and inventions demand collaboration across various kinds of networks. More and more collaborative efforts, instead of individual hero-inventors, resulted in the innovations in the last two centuries and, even more apparently, today. This project has focused on a community of scientists in one large institution instead of individual scientists. The project intended to use comprehensive data from cross disciplines and perspectives to discover meaningful patterns in the history of innovation. The presentation will discuss various research methodologies applied to different types of data used by a research team consisting of more than 10 faculty members and research assistants from the disciplines of information science, history, geography, physics, visual communication design, and mass communication. The presenters will share the integrated research findings regarding the sophisticated relationships and networks of contributed factors and impacts over LCI's 50-year history that complement traditional study of the history of science and technology. The presentation also aims to share our lessons and roadmaps of taking smart data approaches with the intention of helping more researchers to overcome the challenges in researching the innovation history in the digital age.

The second presentation will be given by Dr. James Lee and Arlene Johnson, co-directors of the Digital Scholarship Center at the University of Cincinnati (UC), along with members of their team. They will describe the team's research entitled "Linked Reading," which uses sophisticated machine logics to allow researchers to directly query, analyze, and visualize or sonify data from multiple independent datasets, including the University of Cincinnati's Elliston Poetry Archive. Since 2010, UC Libraries and the Department of English & Comparative Literature have collaborated on The Elliston Project, an audio archive of over 700 recordings of poetry or poetry-related content. The recordings span seven decades and include over 450 poets, including Wendell Berry, Robert Frost, Allen Ginsberg, Louise Glück and a host of others. Alan Liu, a pioneer of digital humanities in literary studies, considers the Elliston project to be a "world-class poetry audio archive," which has the potential to "alter the dominant understandings of a 'digital archive' developed for textual materials." Linked Reading allows one to examine scholarly questions on the question of poetry from multiple angles at once by pivoting laterally between multiple audio and text datasets. A unique opportunity for researchers and educators lies in constellations of micro and nano datasets that have been inadequately studied or even ignored. This approach, gathering smaller datasets of creative materials into a linked network, allows the team to leverage the local strengths in the humanities and creative arts (represented in stellar fashion by Elliston) to facilitate heretofore impossible research projects. Consider, for instance, poetic tone. It's a basic tenet of poetry instruction that the poem on the page is but a score to be performed; and yet, poetry scholarship is in virtually every instance a study of the printed poem. The project illustrates a potential of reshaping this well-studied topic by 'reading' the sonic features of poetic tone in massive numbers of poetry recordings across many linked smaller archives. As such, the aims of this research are: 1, to transform the techniques of linked data into an analytical and interpretive method, and 2, to adapt well-establish machine learning techniques honed on text datasets for the analysis of large archives of born-audio creative works.

Bibliography

Bos, P. (2015). Impact of our graduates on the industry. In: Morgan, S. et al. (eds.) 50 Years of Innovation, pp.34-35. Kent, OH.: Kent State University.

Kobielus, J. (2016). The Evolution of Big Data to Smart Data. Keynote at Smart Data Online 2016 July 13.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the humanities. Journal for Digital Humanities. 2(3) http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Long Papers

Digital Humanities as Critical University Studies

Matt Applegate
mapplega@gmail.com
Molloy College, United States of America

In her 2011 article, "All the Digital Humanists are White, All the Nerds are Men, but Some of Us Are Brave," Moya Z. Bailey unearths a radical, unrealized potential in the Digital Humanities that few have made explicit since its articulation: a practice of the digital humanities that is also a practice of critical university studies. I cite Bailey here in full:

> In blog posts, Miriam Posner and Bethany Nowviskie have both addressed the structures that impede women from connecting to digital humanities. The increase of women in higher level positions within universities have led to changes in the infrastructure, with child care and nursing nests cropping up on campuses across the country. Similarly, people of color have been engaging in critical university studies long before the 1990s when the field is said to have emerged. By demanding space as students and faculty, in addition to advocating for rights as the laborers that built and maintain these institutions, people of color have organized through concerted effort to bring about changes in institutional culture and structure.

The question of diversity and inclusion that informs this intervention is intimately linked to institutional austerity and the precaritization of intellectual labor. Post-2008, the economic pillaging of the university is undeniable. It is also undeniable that the digital humanities emerged as a contemporary force for disciplinary transformation during this precise economic shift. Matthew K. Gold states this matter of factly in his introduction to the 2012 *Debates in the Digital Humanities* anthology, "The Digital Humanities Moment": "At a time when many academic institutions are facing austerity budgets, department closings, and staffing shortages, the digital humanities experienced a banner year that saw cluster hires at multiple universities, the establishment of new digital humanities centers and initiatives across the globe, and multimillion-dollar grants distributed by federal agencies and charitable foundations."

The university's loss coextensive with DH's boon requires further interrogation. This paper builds on Bailey and Gold's work by directly linking concerns in DH for diversity and inclusion to economic disparities in the university via critical university studies. I do so not to condemn DH for its rise, but to continue to unearth its radical potential. While the literature in critical university studies is broad, I connect DH to two critical university approaches in particular: decolonial feminism and Autonomist Marxism. Both approaches augment current debates in DH as they forefront questions of inclusion, diversity, and economic variance, but also provide more pointedly political approaches to pedagogy and tool-use.

In her 2003 book, *Feminism Without Borders*, Chandra Talapade Mohanty claims that "the moment we tie university-based research to economic development–and describe this research as fundamentally drive by market forces–it becomes possible to locate the university as an important player in capitalist rule" (173). This claim is couched in a decolonial method committed to developing "the urgent political necessity of forming strategic coalitions across class, race, and national boundaries," but it is also motivated by a commitment to feminist struggle (9). The university is a site of decolonial feminist struggle in particular because it is a "contradictory place where knowledges are colonized but also contested [...] It is one of the few remaining spaces in a rapidly privatized world that offers some semblance of a public arena for dialogue, engagement, and visioning of democracy and justice" (170). What follows is therefore a simple claim, but one that is difficult to reconcile in a contemporary context, especially as it might apply to DH: "Feminist literacy necessitates learning to see (and theorize) differently–to identify and challenge the politics of knowledge that naturalizes global capitalism and business-as-usual in North American higher education" (171).

It does not take a careful reader to detect a radical undercurrent to Mohanty's interest in feminist literacy, nor should it be a surprise that those disproportionately affected by institutional inequity might rely on a radical political logic with which to situate their intellectual labor. Perhaps the strongest emergent DH interest in which Mohanty's work carries the most methodological weight, however, is found in Roopika Risam's essay, "Navigating the Global Digital Humanities: Insights from Black Feminism." There, Risam argues that

> As the field of digital humanities has grown in size and scope, the question of how to navigate a scholarly community that is diverse in geography, language, and participant demographics has become pressing. An increasing number of initiatives have sought to address these concerns, both in scholarship–as in work on postcolonial digital humanities or #transformDH–and through new organizational structures like the ALliance of Digital Humanities Organizations (ADHO) Multi-Lingualism and Multi-Culturalism Committee and Global Outlook::Digital Humanities (GO::DH), a special interest group of ADHO.

We see similar issues at work in #transformDH and feministDH more broadly. However, Alan Liu's recent claim to a critical infrastructure studies augments these concerns. Liu summarizes his interest in critical infrastructure studies as a “call for digital humanities research and development informed by, and able to influence, the way scholarship, teaching, administration, support services, labor practices, and even development and investment strategies in higher education intersect with society.” The rhetorical shift from “critical university” to “critical infrastructure” is interesting here. Where Liu goes so far to say that most, *if not the whole of our lives*, are organized through institutional mechanisms formative of a “social-cum-technological milieu,” “the word ‘infrastructure’ give[s] us the same kind of general purchase on social complexity that Stuart Hall, Raymond Williams, and others sought when they reached for their all-purpose word, ‘culture.’” Paired with Risam’s work above, Liu draws us to closer to a critique that would mirror Mohanty’s.

At the same time, Mohanty’s decolonial approach dialogues with Autonomist Marxist approaches to the same problem. Writing of their work with CAFA (Committee for Academic Freedom in Africa), George Caffentzis and Silvia Federici comment on institutional formations like those that Mohanty invokes, but also those that are already operative in DH: global initiatives organized around a common goal. Where Caffentzis and Federici depart from the question of DH infrastructure is certainly a question of technological focus, but also it is also a political one. “As was the factory,” Caffentzis and Federici write, “so now is the university” (125). The import of this claim comments on our institutional alliances, as well as our collective understanding of what educational institutions are for. Thinkers of critical university studies define the university this way because it maximizes the forms of solidarity that are available to us in the face of sovereign institutional control.

For both DH and the Autonomist approach, solidarity is most prominently featured in tool-use and production. Following Caffentzis and Federici, Gigi Roggero mobilizes critical university studies toward a reinvention of the tool. In his article, “Notes on Framing and Reinventing Co-research,” he argues that “tools of inquiry have to be reinvented at the level of the general intellect’s networks, going beyond the division between the virtual and the real,” in order to maximize living labor’s break with capital, opening up a space for co-research to form a “material base for revolution” (520-521). DH’s reinvention of the library, the archive, and the application of technology to humanistic inquiry more generally have never been more apt. At the same time, a strong dialogue with Liu and Risam’s work, stemming from Bailey’s claim to DH as critical university studies, is brought to the fore in Roggero’s work.

This paper concludes by theorizing what forms of alliance/solidarity might be drawn between DH’s transformative work at the level of infrastructure with critical university studies’ political work at the level of the institution. I argue that the university is not a freestanding institution; it is embedded within processes of real subsumption that span the whole of contemporary life. Concerns for diversity and inclusion are contoured by this fact, and the transformative power of tool-use extant in DH praxis resist it.

Machine Vision algorithms on cadaster plans

Sofia Ares Oliveira
sofia.oliveiraares@epfl.ch
École Polytechnique Fédérale de Lausanne, Switzerland

Frederic Kaplan
frederic.kaplan@epfl.ch,
École Polytechnique Fédérale de Lausanne, Switzerland

Isabella di Lenardo
isabella.dilenardo@epfl.ch
École Polytechnique Fédérale de Lausanne, Switzerland

Introduction

Cadaster plans are cornerstones for reconstructing dense representations of the history of the city (di Lenardo and Kaplan, 2015). They provide information about the city urban shape, enabling to reconstruct footprints of most important urban components (buildings, streets, canals, bridges) as well as information about the urban population and city functions (census information, property, rent prices, etc.) (Noizet et al, 2013). Cadaster plans are usually the results of coordinated campaigns with standardized methods of measurement and representation. This means that large sets of documents follow the same representation conventions. This regularity opens the possibly of efficient automated process for analyzing them and possibly transforming the information they contain in geo-referenced databases that can be used as part of historical geographical information systems (Gregory et al, 2001).

However, as some of these handwritten documents are more than 200 years old, the establishment of a processing pipeline for interpreting them remains extremely challenging. This may explain why, to our knowledge, no such system exists in the literature. This article reports our effort in this domain, presenting the first implementation of a fully automated process capable of segmenting and interpreting Napoleonic Cadaster Maps of the Veneto Region dating from the beginning of the 19th century. Our system extracts the geometry of each of the drawn parcels, and classifies, reads and interprets the handwritten labels. We believe the general principle of technologies used in the process could be adapted to other cadastral funds, although this has not been tested in the present study.

Methodology

Literature on map processing includes works on many different types of maps, from roads to topographic maps, including hydrographic and cadastral maps. Most studies focus on particular problems and features and thus develop techniques that are highly map specific (Chiang et al, 2014).

Our work addresses the particular case of the Napoleonic cadaster of Venice dated 1808, but aims at developing a method highly adaptable to other cadasters with little extra effort.

We propose a system that segments the cadastral map, identifies and extract segmented objects such as parcels and identifiers and recognizes the extracted hand-written digits. A demo code with examples of the results can be found on Github.
The method is summarized in Fig. 1.

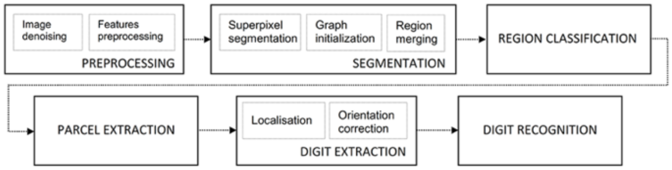


Figure 1: Overview of the system

Preprocessing

Usually, the processed images are ancient documents that have been digitized. To deal with the natural ageing of paper and eventual spots on the map without losing details, we use a non-local means de-noising method (Buades et al, 2005) to smooth the image.

Segmentation

We address the task of extracting the desired information from the document as a segmentation problem, which is a recurrent problem in image processing. A graph-based segmentation approach is adopted, which models the image as a weighted undirected graph. This allows to process the pixels or regions in the spatial domain of the image but also to use higher level information such as connections, similarities and dependencies between the elements.

Because a group of pixels sharing some similarities are more perceptually meaningful than a simple pixel, we use SLIC method (Achanta et al, 2012) to create superpixels. Superpixels are clusters of pixels that share similarities and spatial proximity and have the advantage of reducing the complexity of image processing tasks.

A graph is a mathematical structure composed of vertices and edges, representing a system of connections or interrelations among a set of objects. It is widely used to model relations, to study information systems or to organize data. In our case, the graph representing the image is initialized with superpixels as vertices. Its edges connect neighboring vertices (superpixels) and each edge has a weight which is a measure of the dissimilarity between neighboring elements. The distance (or dissimilarity) metric is based on color and edge/ridge features.

The oversegmentation of the image resulting from superpixel generation is then reduced by grouping superpixels into homogeneous regions and merging the corresponding graph vertices. Our approach uses global homogeneity, meaning that the method minimizes intragroup dissimilarity and maximizes intergroup dissimilarity. The 'dispersion' of edge weights (i.e standard deviation within a region) allows to spot high-weighted edges within a group and thus disconnect dissimilar vertices (i.e remove their edge) to end up with independent homogeneous regions.

Region Classification

The merged regions are classified into 3 classes: text, contour/delimitations and background (smooth textures such as parcels or streets) using a SVM classifier. The training data is composed of manually annotated samples of maps coming from the Napoleonic cadaster of Venice.

Parcel Extraction

The classification results allow the determination possible parcels candidates. A flood fill algorithm is applied, using a ridge detector to indicate boundaries. The chosen ridge detector was originally developed as a vessel enhancement filter (Frangi et al, 1998) and looks for multiscale second order local structures of the image that can be considered tubular. The obtained measure indicates how similar the structure is to a tube, and so it is able to detect ridges. Starting from one point in the regions labeled as background (seed point), the flood fill algorithm floods each zone, i.e parcels, streets, etc. and stops at the boundaries (output of the ridge detector).

Each parcel of the image is extracted as a polygonal shape and the polygon's corner points are stored in GeoJSON format. If the image file is geo-referenced and contains geographical information (a GTIFF file for instance), polygons are exported according to the spatial reference system provided. This allows a fast and easy integration of the shapes into a geographic information system (GIS) and geographic information on the parcels can easily be collected.

Digit Extraction

The parcel identifier is usually contained within the parcel. This observation and the extracted polygons' information can be used to correct misclassified text regions and improve identifier extraction. Elements labeled as text regions are localized, delimited by bounding boxes and grouped so that neighboring characters are extracted together. Again, information from polygons is used to determine whether neighboring digits belong to the same identifier or not (i.e whether neighboring digits are located in the same parcel/polygon). Boxes that do not correspond to identifiers or digits are removed according to specific criteria. Finally, the boxes containing the parcels' identifiers are extracted.

Since the digit recognition step requires horizontally oriented digits to output accurate prediction, the identifiers' boxes are rotated. A principal analysis component is applied to the binary image of the extracted numbers to determine the angle of the rotation.

Digit Recognition

The horizontally oriented numbers are separated into digits that are processed individually. A good digit segmentation is primordial since connected or overlapping digits lead to incorrect recognition. A Convolutional Neural Network (CNN) with two convolutional layers, two fully connected layers and a final softmax layer for multiclass classification is used to predict the identifiers. The CNN is trained on a mixed dataset composed of MNIST dataset (LeCun et al, 1998) and digit samples from Sommarioni register and has a performance of 99.1%. When predicting the numbers, the network outputs the inferred number with a confidence level indicating the reliability of the result.

Results

The proposed approach shows promising results in parcel extraction and identifier recognition. We performed the first 'proof-of-concept' evaluations on manually labeled data taken from different cadaster samples. The total number of annotated objects are shown in Table 1.

Most parcels and identifiers were correctly extracted (Table 2 & 3), which assured us of the feasibility of their automatic extraction. The precision can still be increased for example by using feedback from digit recognition results, i.e, the prediction and its confidence level permit the discarding regions where no reliable identifier has been recognized.

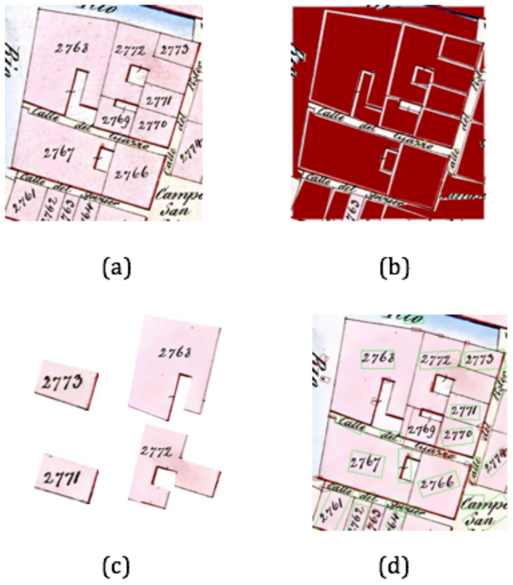


Figure 2: Sample of results: (a) original image, (b) polygon approximation of parcels, (c) extracted parcels and (d) identifier localization

Parcels with labels	810
All parcels (with and without labels)	1185
Parcels' numbers	736

Table 1: Count of ground-truth objects

	Labelled parcels			All parcels		
IoU	> 0.6	> 0.7	> 0.8	> 0.6	> 0.7	> 0.8
Recall	0.77	0.76	0.72	0.72	0.69	0.60
Precision	0.55	0.54	0.51	0.75	0.71	0.62
Ground-truth	810			1185		
Total extracted	1144					

Table 2: Results of parcel extraction with different Intersection over Union (IoU) thresholds

Concerning the digit recognition, only 10% of the identifiers had their digits correctly recognized. Since the models used have shown good performance on nicely detached digits, this is not the fault of the recognition algorithm itself but rather of the digit segmentation procedure. The current segmentation is the main hindrance to an efficient digit recognition, thus, further work should focus on a better number processing algorithm. Another alternative is to avoid the segmentation problem and use a recurrent neural network such as LTSM to process the number as a sequence.

Perspectives

Our work shows promising results for easing and accelerating cadaster processing, especially given our method's efficient parcel segmentation and digit identification. Moreover, the export of a parcel's geometry into GeoJSON format opens up further perspectives to efficiently geo-reference ancient maps. The system can be extended and integrated into a user interface to take better advantage from the results, for example by allowing the user to correct or add information about parcels and identifiers.

Inter	> 0.5	> 0.7	> 0.9
Recall	0.90	0.87	0.81
Precision	0.58	0.55	0.51
Ground-truth	736		
Total localized	1152		

Table 3: Results of parcels' number localization with different Intersection (overlapping percentage) thresholds.

	Correct number	Partial number			
		4 digits	3 digits	2 digits	1 digit
MNIST	58 (.09)	17 (.03)	105 (.16)	94 (.14)	165 (0.25)
MNIST-Sommarioni	66 (.10)	20 (.03)	90 (.14)	103 (.16)	163 (.26)
Total localized	637				

Table 4: Results of parcels' number recognition

The proposed method creates a bridge between previously seperate two data types: the raster object and the vector object. Currently, web-mapping tools consider vector objects as separate layers on the raster maps, and each object needs to be manually redesigned. The automatic vectorization process enables us to perform the visualization and annotation processes directly on the cartographic source without the prerequisite of complex skills. It should greatly facilitate large scale exploitation of such kinds of documents.

Bibliography

di Lenardo, I. and Kaplan, F. (2015) “Venice time machine: Recreating the density of the past,” in Digital Humanities 2015, no. EPFL-CONF-214895.

Noizet, H., Bove, B., and Costa, L. (2013) “Paris de parcelles en pixels.”

Gregory, I. N., Kemp, K. K., and Mostern, R. (2001). “Geographical information and historical research: Current progress and future directions,” History and Computing, vol. 13, no. 1, pp. 7–23.

Chiang, Y.-Y., Leyk, S., and Knoblock, C. A. (2014). “A survey of digital map processing techniques,” ACM Computing Surveys (CSUR), vol. 47, no. 1, p. 1, 2014.

Buades, A., Coll, B., and Morel, J.-M. (2005). “A non-local algorithm for image denoising,” in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 2, pp. 60–65, IEEE.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2012). “Slic superpixels com- pared to state-of-the-art superpixel methods,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 11, pp. 2274–2282.

Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). “Multiscale vessel enhancement filtering,” in Medical Image Computing and Computer-Assisted Interventation—MICCAI’98, pp. 130–137, Springer.

LeCun, Y., Cortes, C., and Burges, C. J. (1998). “The mnist database of handwritten digits”.

Strathy, N.W., C. Y. Suen, C.Y., and Krzyzak. A. (1993). “Segmentation of handwritten digits using contour features,” in Document Analysis and Recognition, 1993., Proceedings of the Second Interna- tional Conference on, pp. 577–580, IEEE

Chen, Y.-K. and Wang, J.-F. (2000). “Segmentation of single-or multiple-touching handwritten numeral string using background and foreground analysis,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, no. 11, pp. 1304–1317

From Usability Testing and Text Analysis to User-Response Criticism

Florentina Armaselu
florentina.armaselu@uni.lu
University of Luxembourg, Luxembourg

Catherine Emma Jones
catherine.jones@uni.lu
University of Luxembourg, Luxembourg

Introduction

Studies on user experience in the digital medium are often related to Human-Computer Interaction (HCI) and the construction of user models or the performance of usability tests in order to support design and evaluation of digital artefacts. User modelling research has mainly focused on the construction of “usable” and “useful” tools providing the users with “experiences fitting their specific background knowledge and objectives” (Fischer, 2001: 65). A variety of characteristics have been used to inform such models, from demographic information (age, gender, native language) or relevant experience (novice, advanced, expert), to interests, goals and plans (general interest categories, task-related objectives/sequences of actions) or contextual information (location, time, physical environment) (Sosnovsky and Dicheva, 2010: 33-34). Many of the approaches merge cognitive science and artificial intelligence (Webb et al., 2001; Biswas and Robinson, 2010; Mohamad and Kouroupetroglou, 2013), whilst usability testing, as a technique from user-centred design, often involves the iterative refinement of a prototype based on user’s feedback (Massanari, 2010). Usability studies also evaluate how a tool is actually used (Brown and Hocutt, 2015) exploring constructs such as ease of use and learnability. Other researches, from the fields of philosophy of technology or digital hermeneutics, go beyond the usefulness and usability aspects of the technology, trying to address questions related to the “human, social, cultural, ethical, and political implications of those technologies” (Fallman, 2007: 296) or to the “self-interpretation of human beings” (Capurro, 2010: 10) in the light of the code. Further directions of study propose a re-orientation of the HCI as “an aesthetic field” (Bertelsen and Pold, 2004: 23) or a cultural perspective on the “reflexive relationship between user and medium” as a “remediation” of the self (Bolter and Grusin, 2000: 230), considered as “humanistic HCI” (Bardzell and Bardzell, 2016).

This article tries to bridge the fields of HCI and Digital Humanities (DH), where HCI techniques are used to evaluate tools developed in DH projects and the results of this evaluation are analysed via DH methods, with the intention of potential development inspired by the literary theory of aesthetic response (Iser, 1980). The paper elaborates on previous work (Armaselu and Jones, 2016) and presents two case studies of usability tests conducted within the framework of interface and game design for digital historical editions and digital cultural heritage (Section 2). Section 3 describes the type of analysis applied to users’ responses, whereby we propose a typology of users and interpretation of users' experience, followed by conclusions and future work (Section 4).

Two case studies

The first case refers to the design and implementation of an XML-TEI-based platform (Transviewer) allowing the exploration of digital editions of historical documents through features for page-by-page navigation, side-by-side view (facsimile/transcription), free-text and named entities search. The usability tests inspired by previous studies (Nielsen, 2000; Jones and Weber, 2012) involved a user-group of 8 researchers in history, political science and linguistics, 4 males and 4 females, aged 25-64. They had to complete 17 tasks using the prototype and to fill-in a USE-based questionnaire (Lund, 2001). During the experiments, the users were asked to think-aloud and the audio and screen interactions were recorded. The common language was English, although none of the participants was a native speaker.

The second case uses data collected during three sessions of gameplay conducted as part of a requirements gathering and co-design process for Pilot 4 of the H2020 Crosscult project. Players were asked to play a board game and contribute reflections as they encountered historical objects pinned to various locations in the city (the Board was derived from a map of Luxembourg City). The first session contained 6 players (5 females, 1 male), the second 5 players (all female) and the third 5 players (4 males, 1 female). All players, aged 25 to 50, worked in a research environment, and none of the participants used English as his/her mother tongue. In the first session, players had 10 roles of the dice to score as many points as possible in successive rounds, the game was shortened so players had to score the most points and reach the end of a score board first.

Analysing user response

For both cases, the users' responses from the questionnaires were transcribed, when not already in electronic form. Partial transcription of the think-aloud audio recordings for the first case was performed (reflections on the experience, improvement suggestions, expressions of disorientation or frustration); the transcription of the second case video-audio recordings is not yet completed, therefore not included in the study. The transcribed snapshots were pre-processed (TXT, XML, R) according to the formats required by the analysis phase. Three types of software were used: Textexture – a tool for representing the text as a network (Paranyushkin, 2011); TXM – a statistical tool for corpus analysis; TheySay – a sentiment analysis package.

The first experiment with Textexture drew attention to noteworthy connections between different clusters of meaning related to users' experience as expressed in their responses. Figure 1 presents two examples: the first highlights how the notion of *trust* is related to the side-by-side view feature of the interface, as allowing the users to compare the transcription with the scanned original and make sure it can be trusted (left); the second illustrates the linking of the sub-networks for *player (reflection, discussion, exchange, opinion*), *place (location, malta, luxembourg)* and *story (card, map, point)*, which reveals the relations, at a conceptual level, between the significant features and interactions of the game.

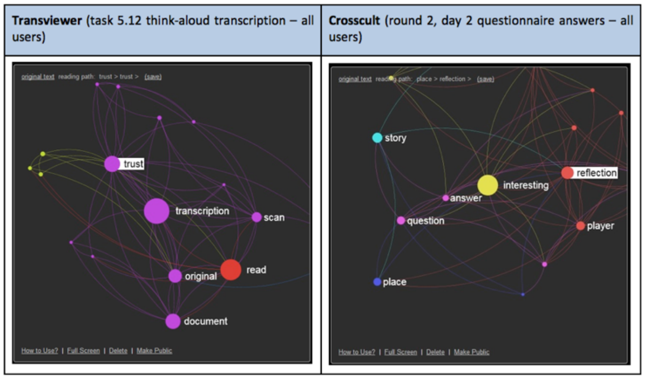

Transviewer (task 5.12 think-aloud transcription – all users)
Crosscult (round 2, day 2 questionnaire answers – all users)
original text reading path: trust > trust > (save)
trust
transcription
scan
original
read
document
How to Use? | Full Screen | Delete | Make Public
original text reading path: place > reflection > (save)
story
interesting
reflection
answer
question
player
place
How to Use? | Full Screen | Delete | Make Public

Figure 1.Textexture

TXM allowed contrasting the specificities scores (Lafon, 1980) corresponding to each user, in terms of overuse/deficit of words usage, as compared to the rest of the corpus. Table 1 shows the positive/negative specificities diagrams based on these measures for three groups of linguistic features. The scores above/under a banality threshold (+/-2.0) indicate highest specificity for responses from particular types of respondents, which allowed us to make hypotheses about a potential "typology" of users that can be described within both case studies.

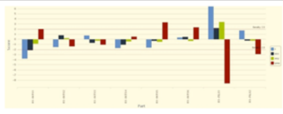
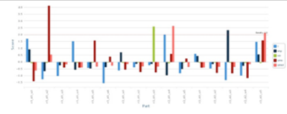
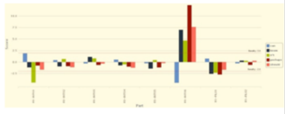
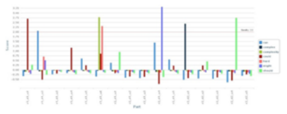
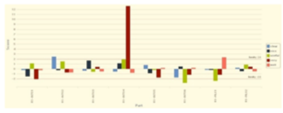
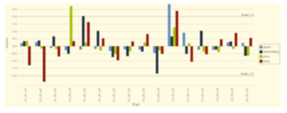
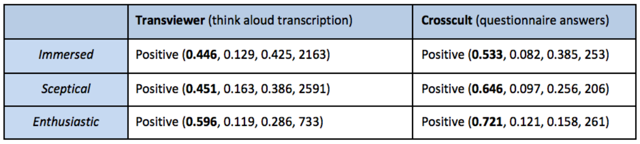
Transviewer	Crosscult
Pronouns: *I, me, my, you*	Pronouns: *I, me, my, us, you, your*
	
Conditionals, ability verbs and negation, uncertainty, difficulty: *can, know, n't, perhaps, should*	Conditionals, ability verbs and negation, uncertainty, difficulty: *can, complex, complexity, could, hard, might, should*
	
Appreciative adjectives, superlatives: *clear, nice, useful, very, well*	Appreciative adjectives, superlatives: *good, interesting, nice, very*
	

Table 1. TXM: User-response specificities

For instance, some users are characterised by an overuse of *I, my* or *you, your*, others by an alternation of them, which can create the impression of an "immersive", "distant" or "versatile" point of view: "Which *I* found strange. Yes, *I* have not yet used the big arrow buttons", "if *you* scroll, *you* have to scroll both" (Transviewer); "prefer to elaborate *my* own answer, without influence", "*I* think it triggers *your* own thinking process" (Crosscult). Similarly, the use of conditionals, negations and uncertainty adverbs are suggestive of a “sceptical” user, in contrast to experiences described with appreciative adjectives and superlatives indicative of an “enthusiastic” standpoint.

After exploring the results in TXM and identifying possible types of users, we analysed the responses via TheySay (Table 2).



	Transviewer (think aloud transcription)	**Crosscult** (questionnaire answers)
Immersed	Positive (**0.446**, 0.129, 0.425, 2163)	Positive (**0.533**, 0.082, 0.385, 253)
Sceptical	Positive (**0.451**, 0.163, 0.386, 2591)	Positive (**0.646**, 0.097, 0.256, 206)
Enthusiastic	Positive (**0.596**, 0.119, 0.286, 733)	Positive (**0.721**, 0.121, 0.158, 261)

Table 2.TheySay: overall and polarity scores (positive, neutral, negative, word count)

The results enabled us to explore differences in sentiment between the types of users. For example, the “enthusiastic” user from both experiments scores highly with respect to the measure of positive polarity, whilst the sceptical user scores are a bit lower but, interestingly enough, higher than the immersed user’s.

It was also observed that sometimes, irrespective the type of user, sentences with high score for humour may actually point to interaction-related aspects like disorientation, confusion, contrariety: “I was ... where was I?”, “I clicked on people but I don't know what happened” (scores 0.996 and 1, Transviewer); “I've never been in the flow because I can't focus on other gamers”, “didn't use any, but I don't think I would“ (scores 0.996 and 1, Crosscult).

Conclusion and future work

The paper describes two case studies in interface and game design dealing with the application of textual analysis to user-response via three systems, for visualisation of the text as a network (Textexture), corpus analysis (TXM), and sentiment analysis (TheySay). The research is still in progress and more experiments with new cases are expected to further support, test and validate the proposed user typologies and interpretation modalities, which might in the future inform humanistic interface design and approaching of user models. In addition, we expect to explore the theoretical matters, assuming that this kind of analysis, beyond its usability-oriented value, may inspire new paths of reflection on user’s self-projection in the digital space, at the intersection of digital hermeneutics, digital aesthetics, and the theory of literary response.

Bibliography

Armaselu, F., Jones, C.E. (2016). “Towards a Digital Hermeneutics? Interpreting the User’s Response to a Visualisation Platform for Historical Documents.” DHBenelux 2016: Conference Abstracts, Belval, Luxembourg. http://www.dhbenelux.org/wp-content/uploads/2016/05/106_Armaselu-Jones_FinalAbstract_DHBenelux_long.pdf.

Bardzell, J., Bardzell, S. (2016). Humanistic HCI. interactions 23, 2 (February 2016), 20-29. DOI=http://dx.doi.org/10.1145/2888576. http://interactions.acm.org/archive/view/march-april-2016/humanistic-hci

Bertelsen, O.W., Pold, S. (2004). “Criticism as an approach to interface aesthetics.” In NordiCHI '04, October 23-27, 2004 Tampere, Finland, ACM 1-58113-857-1/04/10, pp. 23-32. http://www.interactiondesign.us/courses/taught/2010_AD590/pdfs/Bertelsen_2004.pdf.

Biswas, P., Robinson, P. (2010). “A brief survey on user modelling in HCI.” In Proceedings of the International Conference on Intelligent Human Computer Interaction (IHCI) 2010. http://www.cl.cam.ac.uk/~pr10/publications/ihci10.pdf.

Bolter, J.D., Grusin, R. (2000). *Remediation: Understanding the New Media,* The MIT Press, 1999, paperback edition 2000.

Brown, M.E., Hocutt, D.L. (2015). “Learning to Use, Useful for Learning: A Usability Study of Google Apps for Education.” *JUS, Journal of Usability Studies*, Vol. 10, Issue 4, August 2015, pp. 160-181. http://uxpajournal.org/usability-study-google-apps-education/.

Capurro, R. (2010). “Digital Hermeneutics: An Outline.” In *AI & Society*, 2010, 35 (1), 35-42. http://www.capurro.de/digitalhermeneutics.html.

Fallman, D. (2007). “Persuade Into What? Why Human-Computer Interaction Needs a Philosophy of Technology.” In *Persuasive* 2007. Yvonne de Kort et al. (Eds.). Heidelberg, Springer, pp. 295-306.

Fischer, G. (2001). “User Modeling in Human–Computer Interaction.” In *User Modeling and User-Adapted Interaction*, 11: 65.doi:10.1023/A:1011145532042, pp 65–86. http://link.springer.com/article/10.1023/A:1011145532042.

Iser, W. (1980). *The Act of Reading: A Theory of Aesthetic Response,* The John Hopkins University Press, 1978, paperbacks edition 1980.

Jones, C., Weber, P. (2012). “Towards Usability Engineering for online Editors of Volunteered Geographic Information: A perspective on learnability.” *Transactions in GIS* 16(4).

Lafon P. (1980). “Sur la variabilité de la fréquence des formes dans un corpus.“ *Mots* N°1, pp. 127-165. http://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008.

Nielsen, J. (2000). “Why You Only Need to Test with 5 Users.” NN/g, Nielsen Norman Group, Evidence-Based User Experience Research, Training, and Consulting. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users.

Massanari, A. L. (2010). “Designing for imaginary friends: information architecture, personas and the politics of user-centered design.” In *New Media & Society*, 12(3) 401–416. DOI: 10.1177/1461444809346722, SAGE. http://nms.sagepub.com/content/12/3/401.full.pdf.

Mohamad, Y., Kouroupetroglou, C. (2013). “User modeling”, Research and Development Working Group Wiki, last modified on 10 May 2013, at 05:07. https://www.w3.org/WAI/RD/wiki/User_modeling.

Paranyushkin, D. (2011). “Identifying the Pathways for Meaning Circulation using Text Network Analysis”, Nodus Labs. Published October 2011, Berlin. http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/.

Sosnovsky, S., Dicheva, D. (2010). “Ontological technologies for user modelling”, Int. J. *Metadata, Semantics and Ontologies,* Vol. 5, No. 1, pp.32–71. http://www.dfki.de/~sosnovsky/papers/IJMSO_5_1_Paper_03.pdf.

USE Questionnaire: Usefulness, Satisfaction, and Ease of use. Based on: **Lund, A.M.** (2001). “Measuring Usability with the USE Questionnaire”, STC Usability SIG Newsletter, 8:2. http://garyperlman.com/quest/quest.cgi?form=USE.

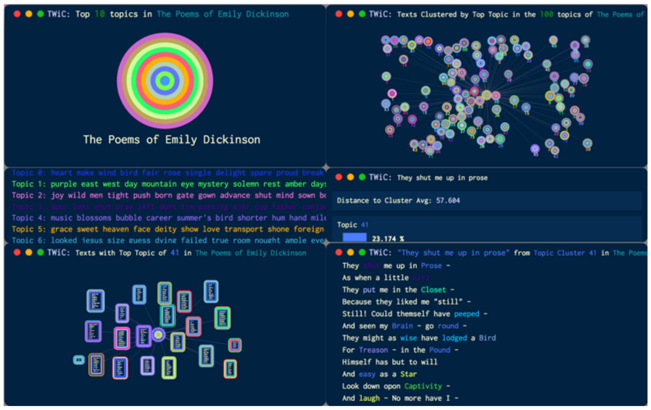
Webb, G.I., Pazzani, M.J., Billsus, D. (2001). “Machine Learning for User Modeling.” In User Modeling and User-Adapted Interaction, 11:19-29, 2001. Kluwer Academic Publishers, Printed in the Netherlands. http://www.umuai.org/anniversary/webb-umuai-2001.pdf.

How to Close Read a Topic Model: TWiC Reads Emily Dickinson's Fascicles

Jonathan Armoza
jarmoza@gmail.com
New York University, United States of America

Abstract

The modeling of mass quantities of cultural objects has led the humanities in new and sometimes discomforting directions. Digital humanists have come to realize the stakes of such practices: emerging paths of scholarship that supplement but may also fundamentally alter the research methodologies and outputs of the humanities. Matthew Wilkens writes that the necessary "decay" of critical techniques that pay close attention to those objects is "a negative in itself only if we mistakenly equate literary and cultural analysis with their current working method" (Wilkens, 2012). One question that remains is if the quantitative working method of such large analyses is compatible with that "current working method" – in this case, the individual interpretation and critique of texts. In the years since Franco Moretti's "distant reading" paradigm became a commonplace, scholars have tested this useful if problematic dichotomy of "close" and "distant" reading. In 2013, for instance, Andrew Piper writes of the "topology" resultant from the dispersive techniques of programmatic text analysis. Setting the lexical components of texts in relation to one another, Piper envisions a tactic of "focalization" that can allow "readers" of such deconstructed texts to understand the relationships between those components (or characteristics) at multiple scales (Piper, 2013: 375). And yet there is always a "between" those multiple scales: "for every unity there can always be something between it and that which it succeeds" (381). Subsequently, there also exist important relationships between these scales of information that inform one another. When macroanalysis brings us to a point where we must return our close attention to our objects of study, we need be reminded of the model that brought us to that perspective in the first place. In focalizing, we are best served to maintain multiple foci. It is a natural tendency to want to confirm macroanalytic results by reading texts and by paying attention to the details of our mathematical models. But how might we do so while keeping the complex relationships of those models in mind?


TWiC: Top 10 topics in The Poems of Emily Dickinson

The Poems of Emily Dickinson

TWiC: Texts Clustered by Top Topic in the 100 topics of The Poems of

Topic 0: heart make wind bird fair rose single delight spare proud break
Topic 1: purple east west day mountain eye mystery solemn rest amber days
Topic 2: joy wild men tight push born gate gown advance shut mind sown bo
Topic 3: door lost shut pray left dont tim passing ajar cup father curiou
Topic 4: music blossoms bubble career summer's bird shorter hum hand mile
Topic 5: grace sweet heaven face deity show love transport shone foreign
Topic 6: looked jesus size guess dying failed true room nought ample eve

TWiC: They shut me up in prose

Distance to Cluster Avg: 57.604

Topic 41
23.174 %

TWiC: Texts with Top Topic of 41 in The Poems of Emily Dickinson

TWiC: "They shut me up in prose" from Topic Cluster 41 in The Poems

They shut me up in Prose –
As when a little Girl
They put me in the Closet –
Because they liked me "still" –
Still! Could themself have peeped –
And seen my Brain – go round –
They might as wise have lodged a Bird
For Treason – in the Pound –
Himself has but to will
And easy as a Star
Look down opon Captivity –
And laugh – No more have I –

Figure 1. “Topic Words in Context”: an in-browser tool for exploring the scales of data in a topic model

When digital humanists use topic models to explore large corpora of texts, they do so at an inherent disadvantage. Typically presented with flat files listing topics and topic weights, they are left to interpret these lists and figures separate from the texts that have just been modeled. Several significant tools have been developed to help scholars visually navigate the textual relationships in topic models. However, in the past few years I have been working on a practical, critical methodology for understanding topic models and the relations between their outputs and that "current working method" of the humanities: human-guided close and contextual reading. For this talk, I will take attendees on a visual exploration of a topic model of Emily Dickinson's poetry using a highly interactive and playful data visualization I developed for my Master's thesis, called "Topic Words in Context" – or "TWiC", for short. TWiC is a multi-paneled environment for web browsers that allows users to explore and juxtapose multiple scales of data in topic models of digital corpora. It uses shapes, colors, and cross-panel highlighting to get users of topic models from "big" data to "small" and back. Importantly, it also provides an alternate "publication" view that resituates modeled texts back into their original publication contexts (e.g. texts split for modeling purposes or texts within a collection). Recalling Piper's topological concepts, TWiC brings our focus simultaneously to these many textual and statistical relationships at play within a topic model. From corpus-wide topic distributions to texts to the topics themselves, each scale of the model when set against each other can reveal hierarchical qualities that enrich and move beyond the semantic/linguistic relationships frequently associated with the word lists of topic models. (Documentation and color screenshots of TWiC are available at in the README.md file at github.com/jarmoza/twic.) Of the many analytical techniques TWiC makes possible, I will demonstrate how we can produce expressive, critical comparisons between our close readings of texts and the smallest of quantitative scales in a topic model: individual texts and individual topics. We will look at different weighting schemes for topic and topic word distributions, how to quantitatively characterize and visualize them, and then how to compare them to traditional literary criticism. As it turns out, the expressiveness of a topic model functions differently depending on the context in which we depict its data. To show this I will turn our attention to the literary-bibliographic focus of my Master's thesis, Emily Dickinson's "fascicle" books of poetry.

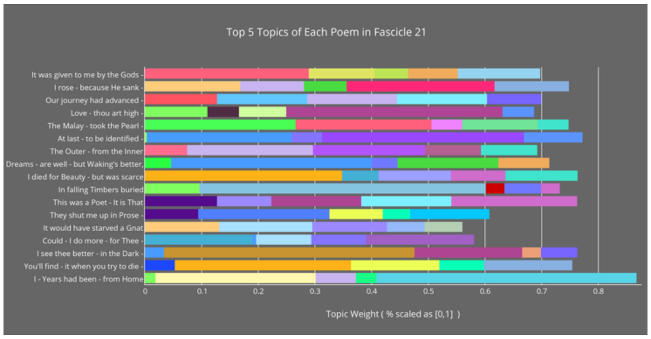
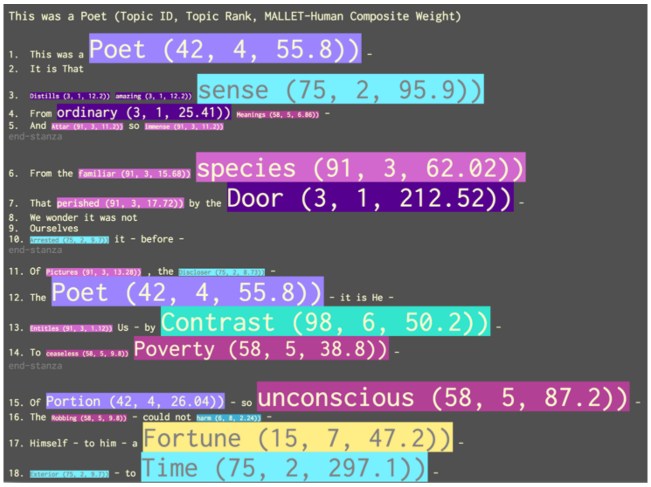


Figure 2. The topics of several Dickinson poems, displayed proportionately and in the order of a fascicle

Emily Dickinson died in 1886, leaving behind in a small, wooden box an unpublished trove of poetry numbering near 2,000 individual works. Many of those poems were bound, hand-sewn into tiny books that have come to be known as her "fascicles." While her poems were being prepared to be published, a family feud arose that split the collection of her manuscripts, and also resulted in the fascicle ordering of her poems being lost for years. A long-studied, now-canonical poet, Dickinson is considered a proto-modernist, someone whose style influenced many of the American poets of the early twentieth century. However, given the size of the task, a comprehensive assessment of every piece of her writing has rarely been attempted. It would not be until the mid-twentieth century that the painstaking effort to rediscover those original orderings was made by bibliographers, notably R.W. Franklin. With that work completed, Dickinson scholars like Eleanor Elson Heginbotham, Sharon Cameron, and others have provided assessments of a select few of these orderings using all of the critical tools at their disposal honed by years of reading Dickinson: interpretations that pay attention to style, context, biography, history, textual studies, and more. Even so it just may not be humanly possible to provide a comprehensive perspective of her writing via such individual attentiveness. Dickinson's poems and their bibliographic history therefore present a fortuitous and somewhat unique set of circumstances for digital humanists. Her oeuvre as a poet is large enough in size to be mathematically modeled. There is a known ordering to much of her works. And her words are "truth told slant" enough in their polysemy to problematize a topic model's expectations of the relationships between them – even at the level of the individual line, let alone across several works.



```
This was a Poet (Topic ID, Topic Rank, MALLET-Human Composite Weight)

1.  This was a Poet (42, 4, 55.8)) -
2.  It is That
3.  Distills (3, 1, 12.2)) amazing (3, 1, 12.2)) sense (75, 2, 95.9))
4.  From ordinary (3, 1, 25.41)) Meanings (58, 5, 6.86)) -
5.  And Attar (91, 3, 11.2)) so immense (91, 3, 11.2))
end-stanza

6.  From the familiar (91, 3, 15.68)) species (91, 3, 62.02))
7.  That perished (91, 3, 17.72)) by the Door (3, 1, 212.52)) -
8.  We wonder it was not
9.  Ourselves
10. Arrested (75, 2, 9.7)) it - before -
end-stanza

11. Of Pictures (91, 3, 13.28)) , the Discloser (75, 2, 8.73)) -
12. The Poet (42, 4, 55.8)) - it is He -
13. Entitles (91, 3, 1.12)) Us - by Contrast (98, 6, 50.2)) -
14. To ceaseless (58, 5, 9.8)) Poverty (58, 5, 38.8)) -
end-stanza

15. Of Portion (42, 4, 26.04)) - so unconscious (58, 5, 87.2)) -
16. The Robbing (58, 5, 9.8)) - could not harm (6, 8, 2.24)) -
17. Himself - to him - a Fortune (15, 7, 47.2)) -
18. Exterior (75, 2, 9.7)) - to Time (75, 2, 297.1)) -
```

Figure 3. A Dickinson poem viewed in a composite weighting scheme that utilizes both topic model weights and the considerations of a literary critic

While digital humanists still want to closely consider our objects of study away from computation, we also want to consider them from these new perspectives that digital modeling methods provide. We tend to bounce between considerations of model-induced order and the contextualizing work done by human-imposed orderings. This talk will provide a case study of how those differing worlds of order relate, how they sometimes either complement or contrast with one another. By combining the outputs of a topic model with the context of Dickinson's fascicle orderings, for instance, one can make quick comparisons between the qualitative discursive relationships of her poems and the quantitative relationships established by a topic model of them. I will introduce TWiC and its publication view to attendees as a means of exploring topic models, showing how it can be used to facilitate close readings of texts that focus on model data as well as on prior humanities criticism of those texts. This exploration will take us from the patterned, colorful shapes of TWiC to several types of analytic visuals that unweave the probabilistic threads of a topic model. By the conclusion, we will be able to proportionally compare the interpretations of a Dickinson critic and a topic model of Dickinson's poems as they are situated in the fascicle books.

Bibliography

Armoza, J. (2014). “Topic Words in Context.” https://github.com/jarmoza/twic.

Armoza, J. (2016). “Topic Words in Context – Dickinson, the Fascicle and the Topic Model.” https://github.com/jarmoza/masters_thesis.

Cameron, S. (1992). *Choosing Not Choosing: Dickinson’s Fascicles*. Chicago, IL: University of Chicago Press.

Dickinson, E. (2013). “Emily Dickinson Archive.” http://www.edickinson.org.

Dickinson, E. (1981). R.W. Franklin (ed), *The Manuscript Books of Emily Dickinson*. Cambridge, MA: Belknap Press.

Franklin, R.W. (1967). *The Editing of Emily Dickinson: A Reconsideration*. Madison, WI: University of Wisconsin Press.

Heginbotham, E. (2003). *Reading the Fascicles of Emily Dickinson: Dwelling in Possibilities*. Columbus, OH: Ohio State University Press.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Mimno, D. (2015). *MALLET: A Machine Learning for Language Toolkit*. https://github.com/mimno/Mallet.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.

Piper, A. (2013). "Reading's Refrain: From Bibliography to Topology." *ELH*. 80(2): 373-399.

Wilkens, M. (2012). "Canons, Close Reading, and the Evolution of Method." In Gold, M. (ed), *Debates in the Digital*

Data Visualization in Archival Finding Aids: A New Paradigm for Access

Anne Bahde
anne.bahde@oregonstate.edu
Oregon State University Libraries and Press
United States of America

Cole Crawford
cole.crawford@oregonstate.edu
Oregon State University, United States of America

Archival finding aids (also called collection guides) are meant to enhance access to archival collections for researchers, and have been presented online for almost two decades. However, researchers still struggle to understand and use them, and the poor functionality of finding aids can often impede the research process. Online finding aids frequently violate many tenets of basic web usability by presenting large blocks of text, complex collection hierarchies, and connections between relevant materials in arcane and unintuitive formats. Many scholars in the humanities have struggled with archival discovery and navigation systems in the course of their research or teaching; cumulatively, these individual annoyances add up to significant interference in the production of humanities scholarship in the humanities.

Ciaran Trace and Andrew Dillon have argued that user problems with archival finding aids may be rooted in the system of power inherent in these tools. According to Trace and Dillon, the traditional archival finding aid has always "reflected, privileged, enabled, and given control to the writer (archivist) more so than to the receiver (researcher)" (Dillon 2012). Users of online finding aids are merely receivers of the information, and in many systems have no control over how the information is presented, or even what information is presented. As a static document, the finding aid limits interpretive possibilities and aggregate analysis, because it cannot be (re)configured to meet users' research needs. Passive, search-based finding aid systems hinder researchers' potential for creativity, and obscure opportunities for serendipitous discovery.

Digital humanities modalities suggest compelling ways to disrupt this power paradigm. Johanna Drucker has called for "knowledge-generating visualizations," which empower the user to produce new interpretations and arguments through manipulation and augmentation of the data presented (Drucker 2014). Tim Sherratt has advocated for the development of digital collection interfaces that enable users to visualize, augment, problematize, and critique collections and collections data (Sherratt 2011). These appeals invite questions about how users might respond to finding aids that present archival information in newly visual ways. What happens for the user when archival finding aids are stretched beyond traditional modalities to invite new interpretations of collections?

This presentation will introduce multiple models for including visualizations in both individual finding aids and discovery systems. Using case studies derived from pilot projects at Oregon State University's Special Collections and Archives Research Center (SCARC), the presenters will discuss models that allow the researcher (and the archivist) to compare, match, recognize, distinguish, arrange, combine, construct, and organize data across a constellation of data points in ways that traditional, textual finding aids cannot. Models discussed will include proof-of-concept designs for interactive geographic timelines, force-directed network graphs, circle packing units, cluster force designs, alluvial diagrams, and treemaps that enable "distant reading" of a finding aid or corpus of finding aids. These visualizations allow users to identify patterns over time or space, relationships between collections, proportions between categories of data, and more. The presenters will explore how these visualizations can impact every stage of the research process, including topic exploration, identification of relevant archival collections, and establishing context and background.

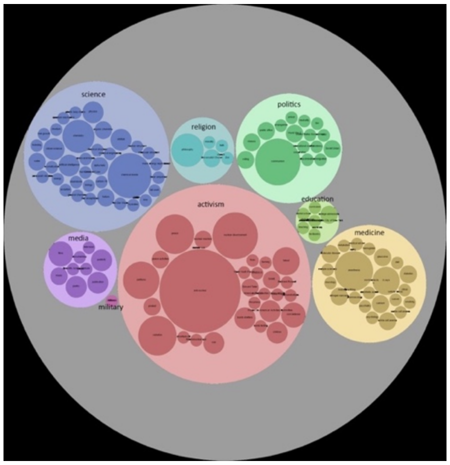


Figure 1: Example of circle-packing of letter keywords from “Global Bonds: Public Activism and Agency in the Letters of Linus Pauling”

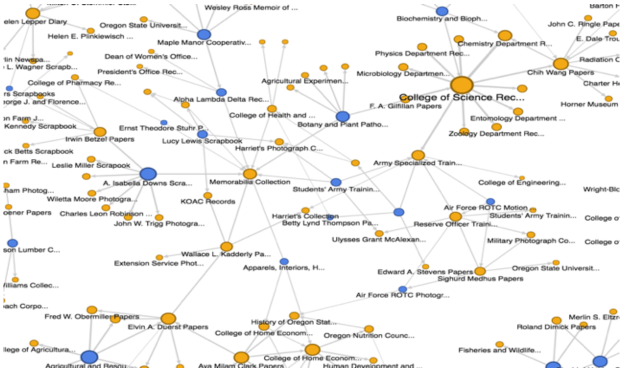

Wesley Ross Memoir of ...
Biochemistry and Bioph...
Oregon State Universit...
Helen E. Pinkiewisch ...
Maple Manor Cooperativ...
Dean of Women's Office...
President's Office Rec...
College of Pharmacy Re...
Alpha Lambda Delta Rec...
College of Health and ...
Botany and Plant Patho...
College of Science Rec...
F. A. Gilfillan Papers
Entomology Department ...
Zoology Department Rec...
Physics Department Rec...
Microbiology Departmen...
Chemistry Department R...
Chih Wang Papers
Horner Museum
Kennedy Scrapbook
Ernst Theodore Stuhr P...
Irwin Betzel Papers
Lucy Lewis Scrapbook
Harriet's Photograph C...
Leslie Miller Scrapbook
A. Isabella Downs Scra...
Memorabilia Collection
Students' Army Trainin...
Army Specialized Train...
College of Engineering...
Wiletta Moore Photogra...
KOAC Records
Charles Leon Robinson
John W. Trigg Photogra...
Harriet's Collection
Betty Lynd Thompson Pa...
Air Force ROTC Motion ...
Students' Army Trainin...
Reserve Officer Traini...
Ulysses Grant McAlexan...
Military Photograph Co...
Wallace L. Kadderly Pa...
Extension Service Phot...
Apparels, Interiors, H...
Edward A. Stevens Papers
Sighurd Medhus Papers
Oregon State Universit...
Air Force ROTC Photogr...
Fred W. Obermiller Papers
Elvin A. Duerst Papers
History of Oregon Stat...
College of Home Econom...
Oregon Nutrition Counc...
Roland Dimick Papers
Merlin S. Eltzr...
Fisheries and Wildlife...
College of Home Econom...
Human Development and ...
Ava Milam Clark Papers
Agricultural and Resou...
Agricultural Experimen...
John C. Ringle Pape...
E. Dale Trou...
Radiation C...
Charter H...
Wright-Bic
College o...

Figure 2: Example of force-directed network graph of related collections

Using several enhanced finding aids designed to promote access to archival materials, the presenters will examine practical considerations for researchers and archivists interested in creating visually enhanced digital finding aids. These considerations include platform selection, tool design, scalability, and data repurposing. SCARC faculty and staff work on multiple platforms, each tailored to specific project needs; choosing the right digital platform is critical to the long-term sustainability of a project, and this choice has only become more difficult with the proliferation of content management systems, digital publishing tools, and repository solutions.

Similarly, selecting the right development frameworks and types of visualizations for a certain data source can either enhance or limit a project’s potentialities. While the presenters have found success with libraries and tools such as d3.js, Leaflet, Sigma.js, and RAW, each archival discovery environment bears its own research complications which necessarily inform tool selection. The presenters will also talk about how both digital products and data production processes scale by considering project management and data entry workflows. Finally, the presenters will address how researchers can incorporate, clean, and repurpose existing data sources into new digital finding aids by examining a legacy finding aid that was repurposed and enhanced.

New research in the humanities often begins with consultation of finding aids and archival research environments. If access to archival collections is obscured by finding aid presentation and form, opportunities for progressive and provocative scholarship are concealed. Redesigned interfaces have the potential to invite new interpretations of collections, to encourage experimentation with research methods, and to plant new seeds of inspiration. Pairing new paradigms of visualization with recently proposed concepts of arrangement and description can open exciting possibilities; rhizomatic organization, for example, interrupts traditional hierarchy, structure, and power paradigms, and suggests a dynamic, acentric finding aid with multiple access points. If archivists create tools to make the archival research process adaptive and dynamic for users with different research needs, humanities scholars are empowered to pursue paths of inquiry never before revealed by traditional access tools.

Bibliography

Drucker, J. (2014). *Graphesis: Visual Forms of Knowledge Production*. Cambridge: Harvard University Press.

Trace, C. and Dillon, A. (2012). “The Evolution of the Finding Aid in the United States: From Physical to Digital Document Genre,” *Archival Science* 12 (4): 501-519.

Sherratt, T. (2011). It’s All About the Stuff: Collections, Interfaces, Power and People.” *Journal of Digital Humanities* 1 (Accessed October 15, 2106) http://journalofdigitalhumanities.org/1-1/its-all-about-the-stuff-by-tim-sherratt/

Aesthetic Appreciation and Spanish Art: Insights from Eye-Tracking

Claire Bailey-Ross
claire.bailey-ross@port.ac.uk
University of Portsmouth, United Kingdom

Andrew Beresford
a.m.beresford@durham.ac.uk
Durham University, United Kingdom

Daniel Smith
daniel.smith2@durham.ac.uk
Durham University, United Kingdom

Claire Warwick
c.l.h.warwick@durham.ac.uk
Durham University, United Kingdom

Introduction

How do people look at and experience art? Which elements of specific artworks do they focus on? Do museum labels have an impact on how people look at artworks? The viewing experience of art is a complex one, involving issues of perception, attention, memory, decision-making, affect, and emotion. Thus, the time it takes and the ways of visually exploring an artwork can inform about its relevance, interestingness, and even its aesthetic appeal. This paper describes a collaborative pilot project focusing on a unique collection of 17th Century Zurbarán paintings. The Jacob cycle at Auckland Castle is the only UK example of a continental collection preserved in situ in purpose-built surroundings. While studies of the psychology of art have focused on individual works and distinctions between representative/non-representative topics, no work has been completed on the aesthetic appreciation of collections or of devotional themes. In this paper, we report upon the novel insights eye-tracking techniques have provided into the unconscious processes of viewing the unique collection of Zurbarán artworks. The purpose of this pilot study was to assess the effects of different written interpretation on the visual exploration of artworks. We will discuss the potential implications of these techniques and our understanding of visual behaviours on museum and gallery practice. The project brings together established research strengths in Spanish art history, experimental psychology, digital humanities, and museum studies to explore, using eye-tracking techniques, aesthetic reactions to digital representations of the individual Zurbarán artworks as well as the significance of the collection as a whole.

Overview

Our experience of art develops from the interaction of several cognitive and affective processes; the beginning of which is a visual scan of the artwork. When regarding an artwork, a viewer gathers information through a series fixations, interspersed by rapid movements of the eye called saccades. The direction of saccades is determined by an interaction between the goals of the observer and the physical properties of the different elements of the scene (e.g. colour, texture, brightness etc). Importantly, studying eye movements offers an insight that does not depend on the participants' beliefs, memories or subjective impressions of the artwork. Previous eye tracking research has highlighted the potential to transform the ways we understand visual processing in the arts (see for example Brieber 2014; Binderman et al., 2005) and at the same time offers a direct way of studying several important factors of a museum visit (Filippini Fantoni et al., 2013; Heidenreich & Turano 2011; Milekic 2010).

Zurbarán's cycle of Jacob and his Sons has been on display in the Long Room at Auckland Castle for over 250 years. It is the only cycle to be preserved in purpose-built surroundings in the UK, and one of very few of its kind in the world. It has a long history in scholarship (Baron & Beresford 2014), but many key aspects of its production and significance have not yet been fully understood. In this study we used eye-tracking in the first stage of exploring audience experience of the extensive Spanish art collections of County Durham, of which the 13 Zurbarán artworks (there are actually only 12 Zurbarán artworks, the 13th Benjamin, is a copy by Arthur Pond) are a key part of, to investigate the ways in which audiences look at Spanish art, how aesthetic experience is evaluated and whether audiences can be encouraged to approach art in different ways. This pilot project primarily investigated how participants visually explore artworks and provides new insights into the potential eye-tracking has to transform the ways we understand visual processing in arts and culture and at the same time offer a direct way of studying several important factors of a museum visit, namely to assess the effects of label characteristics on visitor visual behaviour.

Method

The aim of this study was to determine whether the accompanying written context influences how digital artworks are visually experienced. Whether contextual information impacts on where participants first look (first fixation), if gallery labels influence the time participants choose to view artworks and, especially, whether it influences their aesthetic appreciation of the works. We expected viewing time for artworks and corresponding labels to be predicted by participants' subjective experiences, artwork related features, and contextual factors. Accordingly, we measured viewing time, fixation, and saccades for each artwork and corresponding label using a fixed eye tracking technology (Tobii TX300) in a laboratory setting.

Forty Six students from the University of Durham participated in this study. All participants had normal or corrected vision, no formal training in arts or art history and received course credit for taking part. All participants gave informed consent. The study was approved by Durham University's Department of Psychology Ethics committee. A third of the participants were randomly assigned to the Museum Context group (nMC =16), who inspected digital images of the paintings in conjunction with the contextualizing labels currently in use at Auckland Castle, which rely heavily on relating the content of individual compositions to the words of Jacob in Genesis 49; a third to the Aesthetic Context group (nAC = 15), who received labels foregrounding issues of aesthetic and interpretive interest; and the final third to the No Context group (nNC = 15), who received only basic attribution data: title of composition, name and date of artist, date of composition, and nature of medium (i.e. "oil on canvas").

All stimuli were taken from Auckland Castle's collection of Jacob and his Twelve Sons by Francisco de Zurbarán. Each participant viewed high-resolution digital reproductions of the original artworks presented on the

Tobii TX300 screen based eyetracker. The stimuli were presented in the same sequence for all participants.

Results

Previous authors have shown that when a human being is portrayed in a painting, gazing behaviour is mostly focused on the human figure, independently of contextual elements also depicted in the image. In particular, attention is given to the face area, and it plays a fundamental role in aesthetic judgment (Ro et al., 2007; Massaro et al., 2012; Villani et al., 2015). Given these considerations, three key regions of interest (ROI) were identified; the head, the clothes and the contextualising element. Saccades and fixations were identified offline in Tobii Studio using the default algorithm (onset/offset criterion of 70 degrees/second and a minimum dwell time of 80ms). The key variables of interest for each ROI were (1) Frequency of First Fixation, (2) Time to First Fixation and (3) Total Fixation Duration.

When comparing the first fixation data across the three participant groups (Museum Context group (MC), Aesthetic Context group (AC) and No Context group (NC)), it possible to see an interesting trend (Fig 1) that suggests that contextual labelling appears to change the proportion of participants fixating on the face. The study revealed that the AC labels succeeded in disbursing the gaze more effectively than those that are current MC labels. In all thirteen paintings, evidence shows that participant visual behaviour changed in response to the written interpretation. This suggests that an aesthetic context labelling approach is more successful in stimulating and/or training the gaze than one rooted in theological extrapolation.

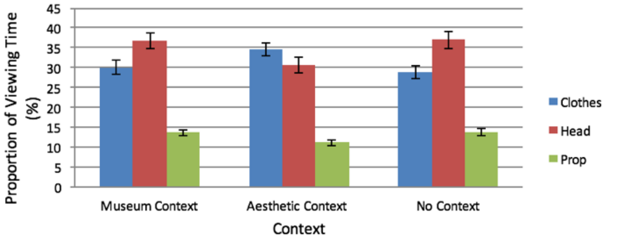


Figure 1: Graph highlighting the proportion of viewing time spent fixating on the face.

The pilot study also found that contextual labelling has a significant effect on influencing levels of aesthetic appreciation, and on the ways in which the gaze can be trained and/or manipulated to engage with areas of interest that would otherwise be overlooked. Reorienting the content of individual labels away from scripture and towards questions of aesthetics and interpretation produced a statistically significant reduction in aesthetic appreciation, which, given that the face is a key driver of aesthetic judgements, is consistent with the finding that the aesthetic context also reduced the participants dwell time on the face.

This paper will also discuss a how participants identify and rank the artworks in terms of authenticity and value. By ranking compositions, we will cross-reference attitudes with the prices (all different) paid by Bishop Trevor at auction in 1756, considering how aesthetic tastes have changed.

Summary and Conclusions

To date, studies of museum and gallery visitor behaviour primarily investigate how people behaviourally and cognitively respond to the design and layout of exhibits. However, they largely ignore the behavioural responses at the 'exhibit- face' (vom Lehn and Heath 2006) or the 'fat moment' (Garfinkel 1967) of visitors' action. Eye-tracking techniques have provided novel insights into the unconscious viewing processes of the 'fat moment' of the unique collection of Zurbarán artworks. The study highlighted statistically significant variations in levels of aesthetic appreciation. More importantly, the experiments indicated that by changing the written interpretation gaze can be redirected towards areas of conceptual significance, challenging the face bias which traditionally plays a fundamental role in aesthetic judgment.

Bibliography

Baron, C., & Beresford, A. (2014). 'Auckland Castle: Zurbarán's Jacob and his Twelve Sons', in Spanish Art in County Durham, ed. Clare Baron & Andy Beresford (Bishop Auckland: Auckland Castle). pp. 26–43.

Bindemann, M., Burton, A.M., Hooge, I.T., Jenkins, R. and De Haan, E.H., (2005). Faces retain Attention. Psychonomic Bulletin and Review, 12(6), pp.1048-1053.

Brieber, D, Nadal, M., Leder, H., & Rosenberg, R. (2014). 'Art in Time and Space: Context Modulates the Relation between Art Experience and Viewing Time', Plos One, 9.6: 1–8.

Filippini Fantoni, S., Jaebker, K., Bauer, D., & Stofer, K. (2013). Capturing Visitors' Gazes: Three Eye Tracking Studies in Museums. In Museums and the Web 2013, Proctor, N., & Cherry, R.,(eds). Silver Spring, MD: Museums and the Web.

Garfinkel, H. (1967). Studies in Ethnomethodology, Englewood Cliffs, N.J: Prentice-Hall.

Heidenreich, S M., & Turano, K.A. (2011). Where Does One Look When Viewing Artwork in a Museum? *Empirical Studies of the Arts,* 29: 51–72.

Massaro, D., Savazzi, F., Di Dio, C., Freedberg, D., Gallese, V., Gilli, G. and Marchetti, A., (2012). When art moves the eyes: a behavioral and eye-tracking study. PloS one, 7(5), p.e37285.

Milekic, S. (2010). Gaze-Tracking and Museums: Current Research and Implications. In J. Trant and D. Bearman (eds). *Museums and the Web 2010: Proceedings.* Toronto: Archives & Museum Informatics.

Ro, T., Friggel, A., & Lavie, N. (2007.) Attentional biases for faces and body parts. Vis. Cogn. 15, 322–348

Villani, D., Morgant, F., Cipresso, P., Ruggi, S., Riva, G., & Gilli, G. (2015). Visual exploration patterns of human figures in action: an eye tracker study with art paintings. Frontiers in Psychology, 6: 1636

vom Lehn, D. and Heath, C. (2006). Interaction at the Exhibit-Face: Video-Based Studies in Museums and Galleries. In H. Knoblauch et al., eds. *Video Analysis: Methodology and Methods. Qualitative Audiovisual Data Analysis in Sociology.* Frankfurt am Main: Lang.

Access to cultural heritage data: a challenge for the Digital Humanities

Anne Baillot
anne.baillot@gmail.com
Centre Marc Bloch, Germany

Marie Puren
marie.puren@inria.fr
INRIA, France

Charles Riondet
charles.riondet@inria.fr
INRIA, France

Laurent Romary
laurent.romary@inria.fr
INRIA, France

Access to Cultural Heritage data is a key issue in the future development of Digital Humanities (Murray-Rust, 2013). Cultural Heritage data encompass digitized resources (scanned artefacts) as well as the attached metadata, annotation or further enrichments, all of which are a necessary basis for reliable computational research. Access to high quality Cultural Heritage data and metadata is, in that sense, the condition for reliable, performing and verifiable research in many arts and humanities fields - and not just of interest to librarians and archivists.

One of the core challenges of giving access to high quality Cultural Heritage data is often the lack of institutional connection between local GLAM Institutions (e.g. Galleries, Libraries, Archives and Museums), infrastructures and research (University College London, 2010; Higgins, 2013). The initiative we want to present in this paper addresses this issue by bringing together several supra-national infrastructures. These research infrastructures develop a common online environment that allows all the relevant actors to connect and improve together access to Cultural Heritage data. This project is based on in-depth exchanges on recognized standards as well as on strategies of access, data curation and management, licensing, and an effort towards open data (Romary et al., 2016).

The "Cultural Heritage Data Reuse Charter" is currently being developed by several organisations and projects grouped together within a steering committee: DARIAH-EU, Europeana, Clarin, E-RIHS and APE together with the European projects Parthenos, HaS and IPERION-CH .

It offers a comprehensive framework regarding all aspects relevant to co-operations revolving around access to and reuse of Cultural Heritage data.

The Cultural Heritage Data Reuse Charter: an encompassing cooperation framework

The Cultural Heritage Reuse Charter is an online environment dedicated to all actors taking part in scholarly reuse of digital data generated by Cultural Heritage Institutions. It addresses five actors: Cultural Heritage Institutions, Cultural Heritage Labs, Researchers, Data Centers and Research Institutions.

- Cultural Heritage Institutions (GLAM) are considered in their function as curators of collections and objects in their physical form and as potential primary initiators of corresponding digital surrogates, from basic descriptions (catalogues of collections, metadata for specific objects) to more elaborate outputs (scans, 3D models, physical analyses, etc.) (Ray, 2014).
- Primary data can be hosted by CHIs or by Higher Education Institutions like universities, but they are in many cases curated by dedicated data centers. These centers play a key role in guaranteeing the stability, the visibility and the long time availability of the primary data. The engagement expected from them in the context of the Charter is of a more technical nature and should ensure a concrete implementation of the CHI-researcher relationship.
- Cultural Heritage laboratories have a high-level expertise in Cultural Heritage. They give essential insights into Cultural Heritage history, technologies, environment, and alteration.
- Researchers are invited to sign in person, independently from the institution for which they are working at the time they sign the Charter. However, academic institutions (departments, universities, research institution or funding agencies) wishing to sign the Charter, or even make it a requirement for their members or the projects they fund or host, are welcome to do so as well.

The Charter environment allows all five actors to declare general principles (common work ethics), and more broadly all the relevant information needed to understand how a given dataset can be reused. It allows its users to get in contact with partners they would want to work with. Institutions can declare their collections; researchers their research interests and existing publications so that these are connected together. Doing so, all of them always have the possibility to define precisely which aspects of their profile information they wish to make public and which not.

By joining forces, and by sharing the information associated to Cultural Heritage collections, the Charter will help document the knowledge generation process and, consequently, increase the quality of data and metadata accessible to research.

Signing the Charter implies making a statement about the technical quality of the data to be reused, or the data derived by such a reuse. The implementation of appropriate standards is considered a key node for the stabilization of data access (Romary, 2011). More broadly, the Charter offers a concrete implementation framework for the FAIR principles (make the data findable, accessible, interoperable and reusable).

The online environment in practice

The principles described below are addressed by a series of components allowing to define the conditions of reuse for each type of data. The Charter environment offers a framework that can be either picked among a set of recommendations or formulated in a text field by the concerned institutions or actors according to their needs and wishes.

This framework encompasses all questions related to the reuse of Cultural Heritage Data:

- Long-term and persistent access to metadata, texts, images (in the case of a manuscript for instance: archival metadata, scan of the manuscript, transcription, annotation)
- Licensing of the content (linking to relevant documentation allowing for instance researchers to gather information on licensing and citation practices they often lack)
- Formats and standards (also connecting to further information)
- Enrichments (connection of scholarly work and CHI work)
- Dissemination of both CHI information and research (visibility of the work of all stakeholders)
- Retro-provision (communicating enrichments based on CHI data to the CHI they originally emanate from)
- Quality control at all levels according to appropriate standards.

In practice, users of the Charter register in the online environment in their primary function as Cultural Heritage Institution, Researcher, Cultural Heritage Lab, Data Center or Research Institution. Identification of entities are realized on the basis of existing standards such as ORCID for researchers.

In the researcher profile, three main areas are to be defined by the registered user. First, he/she has to abide to the reuse principles defined by the Cultural Heritage Institutions regarding the collection he/she wants to work on; this is the “use of primary data” area. Second, he/she has to declare the dissemination principles he/she favours. In this “dissemination of secondary data” area, he/she can gather information on licences. The third area is that of the “cooperation ethics”, in which the researcher declares that he/she will follow best practices in citing the other Charter partners involved in his/her endeavour. This threefold profile is the basis on which the researcher can reach out to institutions or collections he/she wishes to work with.

Timeframe

The Cultural Heritage Reuse Charter is currently under development. Workshops in which information will be gathered especially on the expectations of Cultural Heritage Institutions will take place in Berlin (November 2016), Paris (November 2016), Rome (January 2017) and Dublin (February 2017). Additional input from the other actors is gathered in parallel (interviews). A soft launch of the web interface is planned for the summer of 2017, so that the interface as well as the benefits for the first signatories can be demonstrated in Montreal .

This abstract is in English in order to reach the widest possible community. Presenting the paper in French or having the slides to the presentation in French would be possible as well.

Bibliography

DeNardis, L. (2011) *Opening Standards: The Global Politics of Interoperability*. Cambridge, Mass.: MIT Press.

Europeana (n.d.) The Europeana Licensing Framework. Accessed October 28, 2016. https://goo.gl/947T4z

Higgins, S. (2013). “Digital Curation. The Challenge Driving Convergence across Memory Institutions”, *The Memory of the World in the Digital age: Digitization and Preservation*, UNESCO. Accessed October 30, 2016. https://goo.gl/JteZQe

Inist (n.d.) « Textes de références », *Libre accès à l’information scientifique et technique*. Accessed October 28, 2016. https://goo.gl/jXtmg4

Murray-Rust, P. (2013) “Open Data in Science”, Serials Review, Vol 34, No 1. Accessed October 28, 2016. https://goo.gl/9ZqdiQ

Osswald, A., and Strathmann, S. (2012), The Role of Libraries in Curation and Preservation of Research Data in Germany: Findings of a survey, IFLA World Congress. Accessed October 30, 2016. https://goo.gl/GSOxRX

Ray, J. (2014), *Putting Museums in the Data Curation Picture*, Springer.

Romary, L. (2011) “Stabilizing knowledge through standards - A perspective for the humanities”, *Going Digital: Evolutionary and Revolutionary Aspects of Digitization*, Science History Publications. <inria-00531019>

Romary, L., Mertens, M., Baillot, A (2016) “Data fluidity in DARIAH – pushing the agenda forward”, BIBLIOTHEK Forschung und Praxis, De Gruyter, 2016, 39 (3), pp.350-357. <hal-01285917v2>

Sabharwal, A. (2015) Digital Curation in the Digital Humanities. *Preserving and Promoting Archival and Special Collections*, Chandos Publishing, 2015.

Suber, P. (2012), *Open access*, The MIT Press, Cambridge, London.

University College London (2010), Advancing Research and Practice in Digital Curation and Publishing. *Summary Report* *and Recommendations of the Workshop on Next Steps in Research, Education and Practice*, 2010. Accessed October 30, 2016. https://goo.gl/hqKuKQ

Toponyms as Entry Points into a Digital Edition: Mapping Die Fackel (1899–1936)

Adrien Barbaresi
adrien.barbaresi@oeaw.ac.at
Austrian Academy of Sciences, Austria

Introduction

The significance of place names exceeds the usually admitted frame of deictic and indexical functions, as they enfold more than a mere reference in space. In the western tradition, a current of reflexion which seems to date back to the 1960s has provided the theoretical foundations of the "spatial turn", whose epitome is the concept of space as emergent rather than existing a priori, and composed of relations rather than structures (Warf, 2009). The emergence of current named "GeoHumanities" (Dear et al., 2011) or "Spatial Humanities" (Bodenhammer et al., 2010), has prompted for a transfer of research objects between disciplines as well as an enforcement of the spatial turn in practice through specific methods of analysis. The common denominator consists in opening up new spaces and experimenting in a transdisciplinary perspective (Domínguez, 2011) in a field which has been evolving at an exponential pace since the last decade (Caquard and Cartwright, 2014).

In this paper, I introduce a visualization of collocations of toponyms in the satirical literary magazine Die Fackel ("The Torch"), originally published and almost entirely written by the satirist and language critic Karl Kraus in Vienna from 1899 to 1936. This work carries heterogeneity at its core and contains a considerable variety of toponyms (Biber, 2001) which are highly significant because of the multinational nature of the Austro-Hungarian empire and the later formation of a territorially diminished state.

In order to provide an additional, synthetic access to a digital edition of the work which is already available online (AAC-Fackel corpus), I set out on a distant reading experiment leading to maps meant to uncover patterns and specificities which are not easily retraceable during close reading. I focus on the concept of visualization, that is on the processes and not on the products (Crampton, 2001), and present them together with a critical apparatus, by giving a theoretical perspective on what is being shown and seen. In fact, digital methods in humanities ought to be criticized (Wulfman, 2014) and the cartographic enterprise bears both a thrill and a risk: "adding more to the world through abstraction", and "adding to the riskiness of cartographic politics by proliferating yet more renders of the world" (Gerlach, 2014).

Extraction of toponyms

The particular task of finding place names in texts is commonly named place names extraction, toponym resolution, or geocoding. A first stage involves the identification of potential geographic references, while a second stage resides in a disambiguation process (Leetaru, 2012). Toponym resolution often relies on named-entity recognition and artificial intelligence (Leidner and Lieberman, 2011). However, knowledge-based methods using fine-grained data – for example from Wikipedia – have already been used with encouraging results (Hu et al., 2014).

The present endeavor grounds on a specially curated gazetteer: during the 20th century there have been significant political changes in Central Europe that have severely affected toponyms, so that geographical databases lack coverage and detail. Consequently, the database developed at the Austrian Academy of Sciences (Academy Corpora) in cooperation with the Berlin-Brandenburg Academy of Sciences (Language Center) focuses on Europe and follows from a combination of approaches: gazetteers are curated in a semi-supervised way to account for historical differences, and current geographical information is used as a fallback. Wikidata API and the Geonames database are used to build the database semi-automatically.

The tokenized files of works to be analyzed are filtered and matched with the database by finite-state automatons (Barbaresi and Biber, 2016): toponyms (single or multi-word expressions) are extracted using a sliding window. A cascade of filters is used: current and historical states; regions, important subparts of states, and regional landscapes; populated places; and geographical features. Disambiguation being a critical component (Leetaru, 2012), an algorithm similar to Pouliquen et al. (2006), who demonstrated that an acceptable precision can be reached that way, guesses the most probable entry based on distance to Vienna (Sinnott, 1984), contextual information (closest-country, last names resolved), and importance (place type, population count). The results are projected on a map of Europe using TileMill.

From collocations to lines of thought

In a further analysis, I visualize co-occurrences of extracted toponyms, which can be considered to be a subset of GeoCollocations (Bubenhofer, 2014), in order to draw sequences, airborne lines following their order of appearance. The word "network" is to be used with circumspection as Latour (1999) suggests. Although it is ubiquitous in the terminology of the spatial turn, the now predominant interpretation in the sense of the World Wide Web suggests an immediacy which is contrary to the acceptions it had before, so that the concept of "meshwork" is more appropriate (Ingold, 2007). I thus interpret Figure 1 as a general meshwork which makes it possible to visualize paths depicting chains of thought (Gedankengänge) as well as their intensity (well-trodden or seldom). If they may reveal spatial patterns that would otherwise remain hidden in texts (Bodenhammer et al., 2010), these linkages are also "mappings and tracing imposed on the data" (Wulfman, 2014) which are not meant to be interpreted without further filtering.

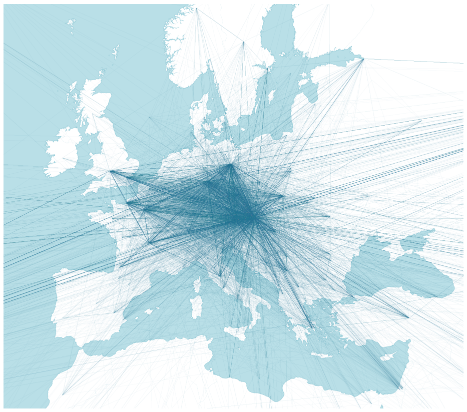


Figure 1. Unfiltered map of toponymic co-occurrences

A rhizome as entry to Die Fackel

That is why I refine the map by selecting a subset of the co-occurrences – the maximal distance between two extracted place names is fixed to twenty tokens – and by color-coding qualitative features – the typology of place names and the axis of time. The most frequent place names are printed out. Surfaces (regions for instance) cannot be represented as such because of historical evolutions and because of the difficulties of linking surfaces without tampering with map readability. Coastlines are depicted in gray to give a sense of orientation, no boundaries are drawn, as they are of a changing nature and may superimpose an artificial reading of the map (Smith 2005).

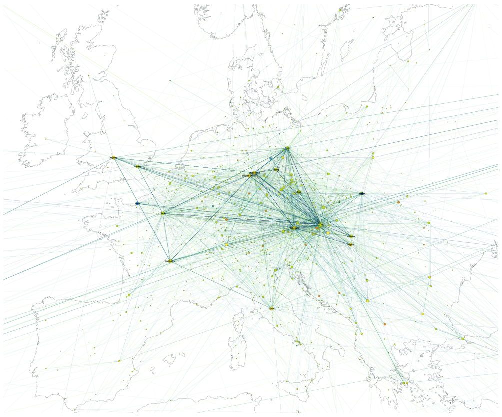


Figure 2. Refined analysis (size proportional to corpus frequency; yellow: sovereign territories; orange: regions; green: populated places; blue: geographical features; time axis represented by a gradient from light green to dark blue)

The notion of rhizome has been used in corpus linguistics by Scharloth et al. (2013) to qualify discourses captured by collocation graphs, it has originally been coined by Deleuze and Guattari (1987 [1980]). This concept is particularly adequate for Kraus, as the Austrian satirist has always been concerned by the multiple aspects of discourse and reality. In addition, his work in Die Fackel evades distant reading processes because of the number of citations used and its ever present and extensive usage of parody. It would be vain to design an authoritative cartography of Die Fackel: following from the principles of heterogeneity and "asignifying rupture" (ibid.), the lines are frequently interrupted. Phenomena in the low-frequency range are filtered out by the human eye, but clusters and interpretation cues may emerge which provide a different access to the work. In this regard, Figure 2 depicts a rhizome connecting heterogeneous information, just as we are all "traversed by lines, geodesics, tropics, and zones marching to different beats and differing in nature" (ibid.).

Conclusion

I have presented a distant reading experiment which consists of connecting toponyms extracted and projected on maps. The latter are meant to be released as an additional feature to the existing digital edition. The Software and gazetteer will be made available under open-source licenses, for development files (please refer to the Github repository). The first example displays unfiltered lines of thought, whereas the second one grounds on a refined analysis and lets the practical image of a rhizome emerge. While Die Fackel criticizes mechanical, instrumental language (Hirt, 2002), the "well-informed" linguistic instruments can help materializing dots or sequences, but not without "human" intervention. The filtering steps on the projection echo the hermeneutic circle of the philological tradition; they also make computed information visible and apprehensible which could remain "blind" otherwise (Barbaresi, 2012).

This is not an authoritative cartography of Die Fackel but rather an indirect depiction of the viewpoint of Kraus and his contemporaries. Drawing on Kraus' vitriolic recording of political life, toponyms in Die Fackel tell a story about the ongoing reconfiguration of Europe. As the map conveys the uncanny sensation that the continent is a field on which points east and west are projected, the lines of force entangle European countries and capitals. Their spatial patterns document an inclination for major cultural centers, whereas the chronological dimension captures a major shift towards the end of publication: the force field intensifies as its range narrows, showing both the interplay of major European powers of the time and the emergence of transatlantic (westwards) and transeuropean (eastwards) relationships. This evolution can be read as an intensification of tensions and a prefiguration of other schemes, this time of military nature.

Bibliography

AAC – Austrian Academy Corpus (2007). AAC–FACKEL, Online Version: “Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936”. http://www.aac.ac.at/fackel

Barbaresi, A. (2012). “La Raison aveugle ? L'époque cybernétique et ses dispositifs”. Conference manuscript: Les critiques de la raison au XXe siècle, University Paris-Est-Créteil, 2012.

Barbaresi, A. and Biber, H. (2016). “Extraction and Visualization of Toponyms in Diachronic Text Corpora”. Digital Humanities 2016: Conference Abstracts, pp. 732–734.

Biber, H. (2001). “In Wien, in Prag und infolgedessen in Berlin – Ortskonstellationen in der 'Fackel'”. Marten-Finnis, S. and Uecker, M. (eds), Berlin-Wien-Prag. Moderne, Minderheiten und Migration in der Zwischenkriegszeit, Peter Lang, pp. 15–26.

Bodenhamer, D. J., Corrigan, J. and Harris, T. M. (2010). The Spatial Humanities: GIS and the Future of Humanities Scholarship. Indiana University Press.

Bubenhofer, N. (2014). “Geokollokationen – Diskurse zu Orten: Visuelle Korpusanalyse”. Mitteilungen des Deutschen Germanistenverbandes, 61(1), pp. 45–89.

Caquard, S. and Cartwright, W. (2014). "”Narrative Cartography: From Mapping Stories to the Narrative of Maps and Mapping”. The Cartographic Journal, 51(2), pp. 101–106.

Crampton, J. W. (2001). “Maps as social constructions: power, communication and visualization”. Progress in Human Geography, 25(2):235–252.

Domínguez, C. (2011). “Literary Geography and Comparative Literature”. CLCWeb: Comparative Literature and Culture, 13(5), 3.

Dear, M. et al. (2011). GeoHumanities: Art, History, Text at the Edge of Place. Routledge.

Deleuze, G. and Guattari, F. (1980). Mille Plateaux. Éditions de Minuit. English translation: 1987, University of Minnesota Press, translation by Brian Massumi.

Foucault, M. (1984). “Of Other Spaces, Heterotopias”. Architecture, Mouvement, Continuité, 5, 46–49. Original Publication: Conférence au Cercle d'études architecturales, March 14th, 1967.

Gerlach, J. (2014). “Lines, contours and legends. Coordinates for vernacular mapping”. Progress in Human Geography, 38(1):22-39.

Hirt, A. (2002). L'Universel reportage et sa magie noire. Karl Kraus, le journal et la philosophie. Kimé.

Hu, Y., Janowicz, K. and Prasad, S. (2014). “Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia”. Proceedings of the 8th Workshop on Geographic Information Retrieval, ACM, pp. 8–16.

Ingold, T. (2007). Lines: A Brief History. Routledge.

Latour, B. (1999). “On recalling ANT”. The Sociological Review, 47(S1):15–25.

Leetaru, K. H. (2012). “Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia”. D-Lib Magazine, 18(9), 5.

Leidner, J. L. and Lieberman, M. D. (2011). “Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language”, SIGSPATIAL Special, 3(2):5–11.

Smith, M. L. (2005). “Networks, territories, and the cartography of ancient states”. Annals of the Association of American Geographers, 95(4), pp. 832-849.

Scharloth, J., Eugster, D. and Bubenhofer, N. (2013). “Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn”. Linguistische Diskursanalyse: neue Perspektiven. Springer Fachmedien Wiesbaden, pp. 345–380.

Warf, B. and Arias, S. (2009). “Introduction: the reinsertion of space into the social sciences and humanities”. The Spatial Turn: Interdisciplinary Perspectives. Routledge, London.

Wulfman, C. E. (2014). “The Plot of the Plot: Graphs and Visualizations”. The Journal of Modern Periodical Studies, 5(1):94–109.

Perseids and Plokamos: Weaving pedagogy, data models and tools for social network annotation

Marie-Claire Beaulieu
marie-claire.beaulieu@tufts.edu
Tufts University, United States of America

Frederik Baumgardt
frederik.baumgardt@tufts.edu
Tufts University, United States of America

Bridget Almas
bridget.almas@tufts.edu
Tufts University, United States of America

The Perseids Project is developing a platform on which students and scholars engage in collaborative acts of scholarship and research on ancient texts (Almas & Beaulieu, 2016). A core value of the project is the focus on pedagogy and the development of undergraduates as researchers. This is complemented by an emphasis on reuse and sharing of tools, data and resources. We keep these values in mind as we develop infrastructure to support complex workflows for the production of new forms of digital publications that are both machine-actionable and human-understandable. In this paper we describe one specific research activity undergraduate students have been conducting on Perseids, the annotation of the social networks of mythological characters. We discuss how opportunities and challenges, both pedagogical and technical, have presented themselves throughout multiple iterations of this effort, and how we evolved the architecture, information structures, and pedagogical workflows in response. We will use our findings to guide future decisions on when to build or reuse tools, and to formulate empirically founded recipes and approaches for specific user scenarios and data types.

The social network annotation project was motivated by an interest in teaching how to produce interpretations of mythological figures and texts. As explained by Schacht, annotation is an activity that is well known to produce deep engagement with a text in the form of close reading while promoting collaboration and conversation among students (Schacht 2016). In this case, we needed to produce interpretations that would be anchored in the primary materials and allow for a representation at the conceptual level. We decided to annotate *Smith's Dictionary of Greek and Roman Biography and Mythology* (Smith's), which offers both a complete narrative for each figure and references to the primary sources on which the narrative is based. This allowed for a double learning outcome. For instance, students would observe that Scylla is directly connected only to first and second generation Titans who represent monstrous or rebellious aspects of nature such as Typhon (volcanoes) and Charybdis (whirlpools). In addition, by following and researching the references to the primary sources, students would note that ancient texts characterize Scylla with words such as "rabid", "aggressive", and "deadly". In this way, students learned that mythological genealogies and social connections are the links which the Greeks made between different concepts represented by the mythological figures. By studying the words with which ancient texts characterize mythological figures, the students understood the value (positive or negative) associated with these concepts in Greek culture.

As we always look to reuse rather than build from scratch when possible, we developed an annotation workflow for this activity using Hypothes.is, an existing open source annotation tool (on the use of Hypothes.is in the classroom, see Kennedy 2016) . We also selected the Standards for Networking Ancient Prosopographies (SNAP) ontology for representation of the social network in the annotations, and the Open Annotation (OA) data model for serialization of the data (Sanderson et al. 2013a) Hypothes.is lacked support for controlled vocabularies, but offered free-form text entry as well as tags, worked on any website, and provided an API for retrieval of the annotations. We prepared explicit tagging instructions for the students with rules that would enable us to apply the controlled terms and data model to the annotations. Students submitted lists of their annotation URIs to the Perseids platform for ingest, review and publication of the data. Perseids software retrieved the students' data from the Hypothes.is API, and upon ingest, applied a transformation, producing OA-compliant annotations using the SNAP ontology. Once the annotations were approved by reviewers in Perseids, we exported the data for final publication via the GapVis interface, to which we added a social network visualization and support for Canonical Text Services data sources.

We completed two full annotation rounds with separate student groups using this workflow. A key finding from a review of the data from the first round was that the lack of ability to visualize the networks at the time of annotation left too much room for error in the directionality of the annotations (On the efficacy of visualization in computer-assisted learning, see Baek and Lane 1988). Despite having explicit instructions on how to identify the subject and object of the annotations, it was difficult for both the students and the reviewers to appreciate their importance without being able to see how their choices impacted the final data. We ended up, for example, with annotations which identified a mother as the son of her child. We tried to address this in the second round by providing even more precise instructions, but the same mistakes were made. Our instructions and transformation rules also became more complex because, having identified the pedagogical significance of the characterizations, we asked students to annotate them as well as the social network connections. Through this process it became clear that we were trying to use the Hypothes.is tool in a way which was very different from the use cases it was developed to support. As a result, we had a workflow which required too much focus on following complex written instructions. This detracted from the pedagogical effectiveness of the activity as well as the overall quality of the resulting annotation data.

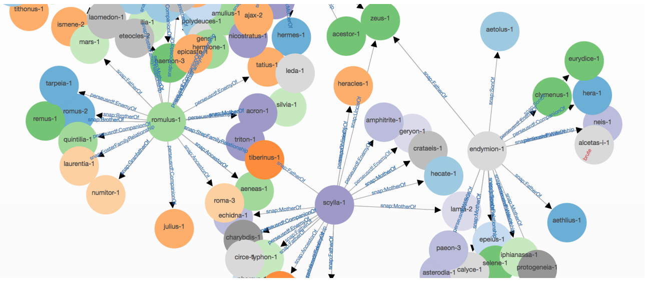


Figure 1. Plokamos network visualization based on students' annotations

At the end of this experimentation phase, we undertook a process of surgical development to address these concerns. With a much clearer understanding of our requirements, and the importance of immediate, visual feedback to the annotation and review process, we developed a targeted interface for semantic annotation which would work on any source text and allow for the data network to be visualized during the annotation process (see Fig. 1). The tool we developed - Plokamos, which is Greek for "something woven" - is a Javascript application backed by an RDF-based triple store. The Plokamos interface is also designed for reuse in other workflows and by other teams. It can be embedded into an existing application and can be extended to support other ontologies and rdf-based annotation bodies. At all times, the data itself remains separate from the tool and available for export and reuse. The configuration is also externalized from the code, and managed, along with the data, as RDF graphs. In our current deployment of Plokamos we reuse Perseids' user model, the Nemo Citable Text Services text browsing interface, and the Apache Marmotta triple store, and we continue use of the OA data model and the SNAP ontology.

We can also see the evolution and objectives of the project reflected in the underlying data structures of the annotations themselves. The annotations consist of a frame with metadata pertaining to their type, provenance and the targeted data source; linked from the frame is the annotation body containing the actual semantics of the annotation. We examine these structures at two architectural levels and from two usage perspectives.

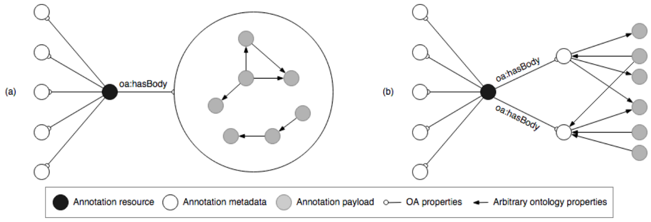


Figure 2. Graph- and Resource-based anchoring of annotation bodies

In designing the body we considered different topologies (of the connection between body and frame; and of the body itself -- structural multiplicity, see Sanderson et al, 2013b) and the compromises they represent between clarity of the annotation body and ease of traversal between annotation frame and body. An annotation body can be embedded into a distinct and uniquely named graph which is referenced by the annotation frame (see Fig. 2 (a)); or it can be anchored through one or more identifying resources which are referenced as the annotation body (see Fig. 2 (b)). The former approach enables quick and easy delineation of individual annotations and allows for complex topologies with multiple graph components. The latter approach offers less flexibility in the structure and complexity of the individual annotations but linking the payload with intermediate resources provides easier pathways to their reuse in other graphs, queries and analyses.

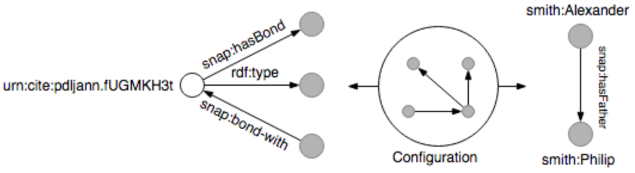
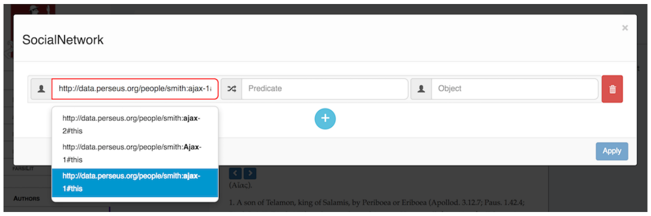


Figure 3. Transformation between machine-actionable and human-readable topologies

The need for the resulting annotation bodies to be understandable by humans as well as algorithmic processing is another factor impacting the data model. Both user groups have different requirements for the topology of the annotation data. Humans may prefer a more direct representation of the data in which object-relational structures are left implicit, while algorithms are not only indifferent to indirect construction but benefit from a more explicit and formal description of the underlying data.


SocialNetwork

http://data.perseus.org/people/smith:ajax-1

http://data.perseus.org/people/smith:ajax-2#this

http://data.perseus.org/people/smith:Ajax-1#this

http://data.perseus.org/people/smith:ajax-1#this

(Αἴας).

1. A son of Telamon, king of Salamis, by Periboea or Eriboea (Apollod. 3.12.7; Paus. 1.42.4;

Figure 4. Annotation interface for entry of social network data

We have used the design of the Plokamos' user interface to help us mediate between these different perspectives. The interface guides the users through the annotation process with a simplified representation of entities and relations in the form of unadorned subject-predicate-object triples (Fig. 4), offering pre-configured choices to to help ensure data integrity, and we use a graph-based system of configuration to transform and expand to the more complex structures in the final annotation data.

Through this iterative approach to supporting the social network annotation activity, putting our core values of pedagogy and reuse front-and-center, we have been able to explore the pedagogical effectiveness of annotation as a learning method with a fairly low initial investment of resources. This allowed us to validate the importance of supporting this activity and refine our understanding of the architecture and data models that would be best suited to it. We were then able to approach the development of custom tools more efficiently, while still designing for maximum extensibility and reuse. The resulting web interface with its RDF-based data source and configuration can be used on a wide variety of existing classroom resources, and expanded upon to support new use cases with varying annotation body, target types, and visual representations.

Bibliography

Almas, B., and Beaulieu, M.-C. (2016) "Scholarship for all!". *Classics Outside the Echo-Chamber: Teaching, Collaboration, Outreach and Public Engagement*, Gabriel Bodard & Matteo Romanello eds. Ubiquity Press. http://www.ubiquitypress.com/site/books/detail/21/digital-classics-outside-the- echo-chamber/

Apache Marmotta (n.d.) - Home,http://marmotta.apache.org/ (accessed 10.25.16).

Baek, Y. K. and Layne, B. H. (1988). Color, graphics and animation in a computer-assisted learning tutorial lesson. *Journal of Computer Based Instruction*, 15(4), 131–135.

Kennedy, M. (2016). Open annotation and close reading the victorian text: Using hypothes.is with students*. Journal of Victorian Culture,* pages 1-9. http://www.tandfonline.com/doi/full/10.1080/13555502.2016.1233905

Nemo. (n.d.) Nemo documentation [WWW Document], URL http://flask-capitains-nemo.readthedocs.io/en/latest/ (accessed 10.25.16).

Perseids Project: Plokamos. (n. d.) perseids-project/plokamos: RDF-based annotations for Perseids, https://github.com/perseids-project/plokamos (accessed 10.25.16).

Rabinowitz, N. (2011). GapVis: Visual Interface for Reading Ancient Texts. Available at http://perseids.org/sites/joth/#index (accessed 9.29.16).

Smith, W. *A Dictionary of Greek and Roman biography and mythology*. London. 1873.

Sanderson, R., Ciccarese, P., Van de Sompel, H. (2013a) W3C. Open Annotation Data Model. Community Draft. 08 February 2013. http://www.openannotation.org/spec/core/

Sanderson, R., Ciccarese, P., and Van de Sompel, H. (2013b) "Designing the W3C Open Annotation Data Model." ACM Press, 2013. 366–375.

Schacht, P. (2016) "Annotation". *Digital Pedagogy in the Humanities*, MLA Commons.

The Hypothesis Project. (n.d.) https://hypothes.is/ (accessed 10.25.16).

SNAP:DRGN. (n.d.) Standards for Networking Ancient Prosopographies. https://snapdrgn.net/ontology (accessed 9.29.16).

Understanding Botnet-driven Blog Spam: Motivations and Methods

Brendon Bevans
brandonbevans@gmail.com
California Polytechnic State University
United States of America

Bruce DeBruhl
bdebruhl@calpoly.edu
California Polytechnic State University
United States of America

Foaad Khosmood
foaad@calpoly.edu
California Polytechnic State University
United States of America

Introduction

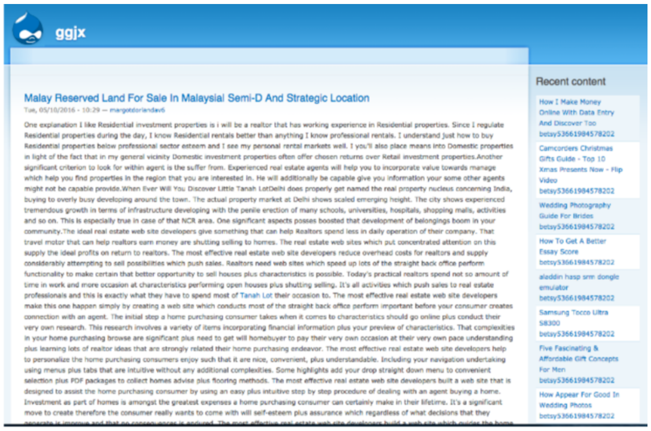
Spam, or unsolicited commercial communication, has evolved from telemarketing schemes to a highly sophisticated and profitable black-market business. Although many users are aware that e-mail spam is prominent, they are less aware of blog spam (Thomason, 2007). Blog spam, also known as forum spam, is spam that is posted to a public or outward facing site. Blog spam can be to accomplish many tasks that email spam is used for like posting links to a malicious executable.

Blog spam can also serve a couple of unique purposes. First, blog spam can influence purchasing decisions by featuring illegitimate advertisements or altering a product's review. Second, blog spam can include content spam with target keywords designed to change the way a search engine identifies pages (Geerthik, 2013). Lastly, blog spam can contain link spam which spams a URL on a victim page to increase the inserted URLs search engine ranking. Overall, blog spam weakens search engines' model of the Internets popularity distribution and leads to malicious pages increasing in popularity. Much academic and industrial effort has been spent to detect, filter, and deter spam (Dinh et al, 2015, Spirin and Han, 2012).

Less effort has been placed in understanding the underlying mechanisms of spambots and botnets. One foundational study in characterizing blog spam (Niu et al, 2007), provided a quantitative analysis of blog spam in 2007. This study showed that blogs in 2007 included incredible amounts of spam but does not try to identify linked behavior that would imply botnet behavior. A later study on blog spam (Stringhini et al, 2015) explores using IPs and usernames to detect botnets but does not characterize the behavior of these botnets. In 2011, a research team (Stone-Gross et al, 2011) infiltrated a botnet which allowed for observations of the logistics around botnet spam campaigns. Overall, our understanding of blog spam generated by botnets is still limited.

Related Work

Various projects have attempted to identify the mechanics, characteristics, and behavior of botnets that control spam. In one important study (Shin et al, 2011), researchers fully evaluated how one of the most popular spam automation programs, XRumer, operates. Another study explored the behavior of botnets across multiple spam campaigns (Thonnard and Dacier, 2011). Others (Pitsillidis et al, 2012) examined the impact that spam datasets had on characterization results. Luzezanu et al. explored the similarities between email spam and blog spam on Twitter (Lumezanu and Feamster, 2012). They show that over 50% of spam links from emails also appeared on Twitter.



ggjx

Malay Reserved Land For Sale In Malaysial Semi-D And Strategic Location

Tue, 05/10/2016 - 10:29 — margotdorlandav6

One explanation I like Residential investment properties is i will be a realtor that has working experience in Residential properties. Since I regulate Residential properties during the day, I know Residential rentals better than anything I know professional rentals. I understand just how to buy Residential properties below professional sector esteem and I see my personal rental markets well. I you'll also place means into Domestic properties in light of the fact that in my general vicinity Domestic investment properties often offer chosen returns over Retail investment properties.Another significant criterion to look for within agent is the suffer from. Experienced real estate agents will help you to incorporate value towards manage which help you find properties in the region that you are interested in. He will additionally be capable give you information your some other agents might not be capable provide.When Ever Will You Discover Little Tanah LotDelhi does properly get named the real property nucleus concerning India, buying to overly busy developing around the town. The actual property market at Delhi shows scaled emerging height. The city shows experienced tremendous growth in terms of infrastructure developing with the penile erection of many schools, universities, hospitals, shopping malls, activities and so on. This is especially true in case of that NCR area. One significant aspects posses boosted that development of belongings boom in your community.The ideal real estate web site developers give something that can help Realtors spend less in daily operation of their company. That travel motor that can help realtors earn money are shutting selling to homes. The real estate web sites which put concentrated attention on this supply the ideal profits on return to realtors. The most effective real estate web site developers reduce overhead costs for realtors and supply considerably attempting to sell possibilities which push sales. Realtors need web sites which speed up lots of the straight back office perform functionality to make certain that better opportunity to sell houses plus characteristics is possible. Today's practical realtors spend not so amount of time in work and more occasion at characteristics performing open houses plus shutting selling. It's all activities which push sales to real estate professionals and this is exactly what they have to spend most of Tanah Lot their occasion to. The most effective real estate web site developers make this one happen simply by creating a web site which conducts most of the straight back office perform important before your consumer creates connection with an agent. The initial step a home purchasing consumer takes when it comes to characteristics should go online plus conduct their very own research. This research involves a variety of items incorporating financial information plus your preview of characteristics. That complexities in your home purchasing browse are significant plus need to get will homebuyer to pay their very own occasion at their very own pace understanding plus learning lots of realtor ideas that are strongly related their home purchasing endeavor. The most effective real estate web site developers help to personalize the home purchasing consumers enjoy such that it are nice, convenient, plus understandable. Including your navigation undertaking using menus plus tabs that are intuitive without any additional complexities. Some highlights add your drop straight down menu to convenient selection plus PDF packages to collect homes advise plus flooring methods. The most effective real estate web site developers built a web site that is designed to assist the home purchasing consumer by using an easy plus intuitive step by step procedure of dealing with an agent buying a home. Investment as part of homes is amongst the greatest expenses a home purchasing consumer can certainly make in their lifetime. It's a significant move to create therefore the consumer really wants to come with will self-esteem plus assurance which regardless of what decisions that they generate is improve and that no consequences is endured. The most effective real estate web site developers build a web site which guides the home

Recent content

How I Make Money Online With Data Entry And Discover Too
betsy53661984578202

Camcorders Christmas Gifts Guide - Top 10 Xmas Presents Now - Flip Video
betsy53661984578202

Wedding Photography Guide For Brides
betsy53661984578202

How To Get A Better Essay Score
betsy53661984578202

aladdin hasp srm dongle emulator
betsy53661984578202

Samsung Tocco Ultra S8300
betsy53661984578202

Five Fascinating & Affordable Gift Concepts For Men
betsy53661984578202

How Appear For Good In Wedding Photos
betsy53661984578202

Figure 1: Browser rendering of the ggjx honeypot.

The underground ecosystem build around the botnet community has been explored (Stone-Gross et al, 2011). In a surprising result, over 95% of pharmaceuticals advertised in spam were handled by a small group of banks (Levchenko et al, 2011). Our work is similar in that we are trying to characterize the botnet ecosystem, focusing on the distribution and classification of certain spam producing botnets.

Experimental Design

In order to classify linguistic similarity and differences in botnets, we implement 3 honeypots to gather samples of blog spam. We configure our honeypots identically using the Drupal content management systems (CMS) as shown in Figure 1. Our honeypots are identical except for the content of their first post and their domain name. Ggjx.org is fashion themed, npcagent.com is sports themed, and gjams.com is pharmaceutical themed. We combine the data collected from Drupal with the Apache server logs to allow for content analysis of data collected over 42 days. To allow botnets time to discover the honeypots, we activate the honeypots at least 6-weeks before data collection.

We generate three tables of content for each honeypot. In the user table, we record the information the spambot enters while registering and user login statistics that we summarize in Table 1. This includes the user id, username, password, date of registration, registration IP, and number of logins. In the content table, we record the content of spam posts and comments which we summarize in Table 2. This includes the blog node id, the author's unique id, the date posted, the number of hits, type of post, title of the post, text of the post, links in the post, language of the post, and a taxonomy of the post from the Alchemy API.

Honeypot	Quantity	Mean Logins/User	# of Countries
ggjx	62992	1.066	83
gjams	28230	1.102	40
npcagent	34332	1.05	53

Table 1: User table characteristics for three honeypots

Honeypot	Quantity	Avg. Hits	Avg. Links	English Posts
ggjx	2279	28.237	2.356	1962
gjams	2225	18.178	0.311	2137
npcagent	1430	29.043	1.823	1409

Table 2: Characteristics for the content tables

Honeypot	ggjx	gjams	npcagent
# Of Entities	3430	1790	1566
# of Users	62992	28230	34332
Mean Users/Entity	18.365	15.771	21.923
Max Users/Entity	37589	14249	23577
σ of Users/Entity	666.128	359.619	611.157
# of IPs	5291	3092	2120
Mean IPs/Entity	1.543	1.727	1.354
Maximum IPs/entity	118	135	60
σ of IP Quantity	4.277	5.551	2.406
Mean Posts/Entity	.664	1.243	.907
Max Posts/Entity	163	484	664
σ of Posts/Entity	5.319	14.448	17.256
% of Entities Who Posted	15.2	12.4	13.5

Table 3: Characteristics of entities

Lastly, in the access table, we include data and meta-data from the Apache logs. This includes the user id, the access IP, the URL, the HTTP request type, the node ID, and an action keyword describing the type of access.

Our honeypots received a total of 1.1 million requests for ggjx, 481 thousand requests for gjams, and 591 thousand requests for npcagent.

Entity Reduction

It is widely accepted that spambot networks, or botnets, are responsible for most spam. Therefore, we algorithmically reduce spam instances into unique entities representing botnets. For each entity, we define 4 attributes: entity id, associated IPs, usernames, and associated user ids. To construct entities we scan through the users and assign each one to an entity as follows.

1. For a user, if an entity exists which contains its username or IP, the user is added to the entity.
2. If more than one entity matches the above criteria, all matching entities are merged.
3. If no entity matches the above criteria, a new entity is created.

We summarize the entity characteristics in Table 3. The maximum number of users in one entity is almost 38 thousand for ggjx with over 100 unique IP addresses. These results confirm what is expected - the vast majority of bots interacting with our honeypots are part of large botnets. This also allows us to perform content analysis exploring what linguistic qualities differentiate botnets.

Feature	Description	Indicates	Effective
Bag or Words	Set of words with count	Lexical content	Yes
Alchemy	Document taxonomy	Taxonomy	Yes
Link	URL core domain names	URL similarity	Variable
Vocab	Vocab complexity	Vocabulary complexity	No
Part-of-speech	A BoW of parts-of-speech	Simple syntax	No

Table 4: NLP feature sets we consider for our content analysis and their effectiveness at differentiating botnets

Content Analysis

To better understand botnets, we use natural language processing (NLP) (Collobert and Weston, 2008) for analyzing the linguistic content of entities. For our analysis, we consider various feature sets as proxies for linguistic characteristics as summarized in Table 4. We use a Maximum Entropy classifier to test which features differentiate botnets. In order to test a feature, we train the classifier with 70% of the posts, randomly selected, from the N largest entities and test it with the remaining 30% of the posts. Our final results are the average of three runs.

The first feature set we test is Bag Of Words (BoW) which models the lexical content of posts. Put simply, each word in a document is put into a 'bag' and the syntactic structure is discarded. For implementation details, see our technical report. In figure 2, we show our analysis of the BoW feature set.

When considering the top 5 contributing entities, the classification accuracy is less than 95% which implies that the lexical content of botnets varies greatly. The second feature we consider, is the taxonomy provided by IBM

Watson's AlchemyAPI [4]. Alchemy's output is a list of taxonomy labels and associated confidences. For the purpose of our analysis, we discard any low or non-confident labels. In Figure 3, we show our analysis of the Alchemy Taxonomy feature set which highlights the accuracy of Alchemy's taxonomy. We note that the Alchemy Taxonomy feature set is dramatically smaller in size than the BoW feature set while still providing high performance. This indicates a full lexical analysis is not necessary but a taxonomic approach is sufficient. Our third feature is based on the links in the posts. To create the feature, we parse each post for any HTTP links and strip the link to its core domain name.

The classifier with the link feature set had varied results, as shown in Table 5, where it was reliable in differentiating ggjx entities but less reliable for the other two honeypots. These results correlate with link scarcity from Table 2.

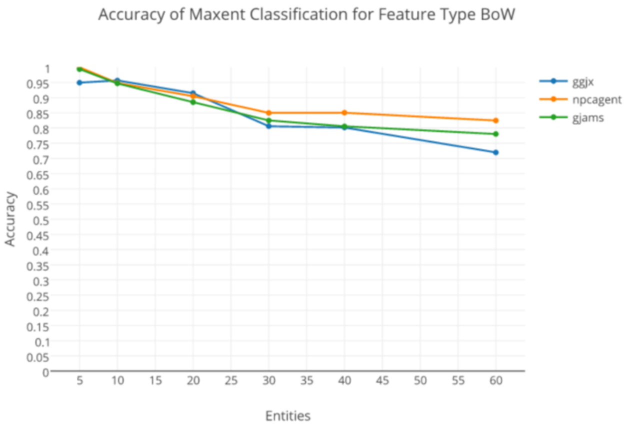


Figure 2

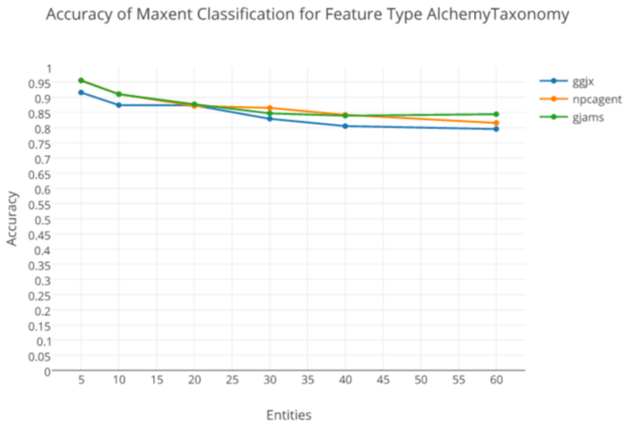


Figure 3

We test the normalized vocabulary size of a post as a feature. We derive this from the number of unique words divided by the total number of words in the post. As shown in Table 5, the vocabulary size does not differentiate botnets.

We also form a feature set based on the part-of-speech (PoS) makeup of a post using the Stanford PoS Tagger. The Stanford PoS tagger returns a pair for each word in the text, the original word and corresponding PoS. We create a BoW from this response that creates an abstract representation of the document's syntax. As shown in Table 5, the PoS does not differentiate botnets.

Feature Set	Database	Accuracy (10 entities)	Accuracy (60 entities)
BoW	ggjx	93%	71%
BoW	gjams	92%	78%
BoW	npcagent	93%	83%
Alchemy	ggjx	87%	80%
Alchemy	gjams	91%	84%
Alchemy	npcagent	91%	82%
Link	ggjx	89%	84%
Link	gjams	53%	37%
Link	npcagent	72%	61%
PoS	ggjx	32%	16%
PoS	gjams	53%	39%
PoS	npcagent	70%	60%
Vocab	ggjx	32%	17%
Vocab	gjams	50%	36%
Vocab	npcagent	74%	60%

Table 5: Accuracies for various features when identifying 10 and 60 entities using the maximum entropy classifier

Conclusions

In this paper, we examine interesting characteristics of spam-generating botnets and release a novel corpus to the community. We find that hundreds of thousands of fake users are created by a small set of botnets and much fewer numbers of them actually post spam. The spam that is posted is highly correlated by subject language to the point where botnets labeled by their network behavior are to a large degree re-discoverable using content classification (Figure 3).

While link and vocabulary analysis can be good differentiators of these botnets, it is the content labeling (provided by Alchemy) that is the best indicator. Our experiment only spans 42 days, thus it's possible the subject specialization is a feature of the campaign rather than the botnet itself.

Bibliography

Apache (n.d.) Apache virtual host documentation, https://httpd.apache.org/docs/current/vhosts/, Accessed: 2016-08-10.

Collobert, R., and Weston, J. (2008) "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, ACM, pp. 160–167.

Dinh, S., Azeb, T., Fortin, F., Mouheb, D., and Debbabi, M. (2015) "Spam campaign detection, analysis, and investigation," *Digital Investigation,* vol. 12, S12–S21.

Geerthik, S. (2013) "Survey on internet spam: Classification and analysis," International Journal of Computer Technology and Applications, vol. 4, no. 3, p. 384, 2013

Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., F´elegyh´azi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., et al.,(2011). "Click trajectories: End-to-end analysis of the spam value chain," in 2011 IEEE Symposium on Security and Privacy, IEEE, 2011, pp. 431–446.

Lumezanu, C., and Feamster, N. (2012) "Observing common spam in twitter and email," in Proceedings of the 2012 ACM conference on Internet measurement conference, ACM, pp. 461–466.

Niu, Y., Chen, H., Hsu, F., Wang, Y.-M., and Ma, M. (2007)“A quantitative study of forum spamming using context-based analysis.,” in NDSS.

Pitsillidis, A., Kanich, C., Voelker, G. M., Levchenko, K., and Savage, S. (2012) “Taster’s choice: A comparative analysis of spam feeds,” in *Proceedings of the 2012 ACM conference on Internet measurement conference,* ACM, 2012, pp. 427–440.

Shin, Y., Gupta, M., and Myers, S. A. (2011). “The nuts and bolts of a forum spam automator.,” in LEET, 2011.

Spirin, N., and Han, J. (2012)“Survey on web spam detection: Principles and algorithms,”ACM SIGKDD *Explorations Newsletter*, vol. 13, no. 2, pp. 50-64.

Stone-Gross, B., Holz, T., Stringhini, G., and Vigna, G. (2011) “The underground economy of spam: A botmaster’s perspective of coordinating large-scale spam campaigns.,” LEET , vol. 11, pp. 4–4

Stringhini, G., Mourlanne, P., Jacob, G., Egele, M., Kruegel, C., and Vigna, G. (2015) “Evilcohort: Detecting communities of malicious accounts on online services,” in 24th USENIX Security Symposium (USENIX Security 15), 2015, pp. 563–578. 8

Thomason, A. (2007)“Blog spam: A review.,” in CEAS, Citeseer.

Thonnard, O., and Dacier, M. (2011) “A strategic analysis of spam botnets operations,” in Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, ACM, pp. 162–171.

Facilitating Fine-grained Open Annotations of Scholarly Sources

Peter Boot
peter.boot@huygens.knaw.nl
Huygens ING, The Netherlands

Ronald Haentjens Dekker
ronald.dekker@huygens.knaw.nl
Huygens ING, The Netherlands

Marijn Koolen
marijn.koolen@huygens.knaw.nl
Huygens ING, The Netherlands

Liliana Melgar
melgar@uva.nl
University of Amsterdam, The Netherlands

Introduction

In the scholarly domain, annotation is a fundamental activity (Unsworth, 2000). Current web-based annotation facilities enable a specific way of annotation (via note-taking, highlighting or commenting) which are useful when scholars are exploring or gathering an initial set of resources, but more sophisticated support is needed for detailed analysis, close reading, and data enrichment. At this point, it is important to take into account the structural relations between documents and their parts. For example, when annotating a letter, annotation tools should be aware that a targeted text fragment is the name of the sender, or that the annotation of a film targets the intellectual work instead of the specific version or copy on which the annotation is made.

In addition, many standalone tools use annotation models with idiosyncratic solutions to enable the relations between different media objects and their parts, which limits the possibilities to exchange those annotations . In general, there is a lack of necessary details for durable access to and interpretation of annotations. For this, detailed information is needed about the annotated object, the annotator and the annotation itself (Melgar et al. 2016, Walkowski & Barker, 2010). In this paper we focus on the requirements for the annotated object, in a web-based environment, and propose a method for making necessary details of objects openly available for any annotation tool.

Requirements of scholarly annotation

In line with (W3C 2017b) we refer to the object that is annotated as the annotation target, the content of the annotation as the annotation body and who or what creates the annotation as the annotation creator. All three are complex entities with aspects that have consequences for interpreting an annotation (Melgar et al., 2016).

Annotation Creator: With respect to the creator it is important to know the intention/motivation for making the annotation (Walkowski & Barker, 2010) and when sharing and reusing annotations, their level of expertise, both in terms of the scholarly domain and in the nature of the annotation task (e.g. the amount of experience/expertise of the annotator in classifying objects according to a controlled vocabulary).

Annotation target: of the target it is important to know which part of the object is targeted. This is not merely about addressing media fragments. Media (e.g., html, mp3, jpg) are carriers of abstract information objects (scenes in movies, chapters in books, objects in pictures) with different conceptual levels (e.g. work, expression or manifestation , see Figure 1) and it is essential to be able to address those abstract objects and the relationships between them.

Annotation body: Of the content of the annotation it is important to know its nature (a natural language comment, structural or subject metadata, a link to another resource), in what form it is made (e.g. closed representation or natural language representation), at what level of control (from mostly uncontrolled to strictly controlled and structured) and for what scholarly purpose, e.g. gathering or exploring sources or thematic or stylistic analysis (Melgar et al., 2017).

State of the Art

There are various models for capturing digital annotations to make them accessible and interpretable. The Web Annotation Data Model (W3C 2017a, 2017b) is a generic model that covers aspects of the annotation body, target and creator. This model focuses on annotations in the context of online social interaction (e.g., commenting, sharing), not necessarily on scholarly annotations done during analysis or data enrichment .

An extended model specifically for scholarly research was proposed by Hunter et al. (2011), which includes context aspects for both the annotation body and target. The Annotating All Knowledge Coalition is also directed at scholarly annotation and lists several issues, including:

1. The lack of support for discovery, sharing and reuse of annotations.
2. Underutilization of collections.
3. The closed and non-standardized nature of current annotation tools.

Current annotation support is either part of a suite of functionalities in monolithic applications with their own models for annotation (e.g. TextGrid , Textual Communities , eLaborate , CATMA for text,Elan and Anvil for multimedia materials, and QDA software packages for mixed media qualitative data analysis), or they lack specificity in describing the annotation target, e.g. Hypothes.is (Perkel, 2015) and Pundit (Grassi et al., 2012) and site-specific annotation tools, e.g. in The Diary of Samuel Pepys).

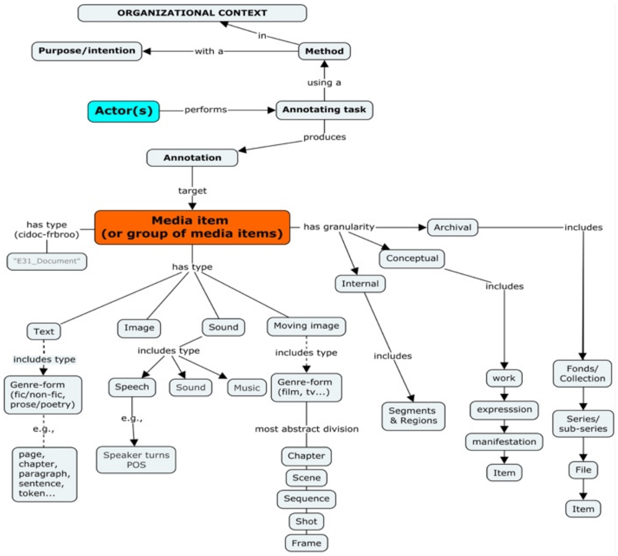


Figure 1. Conceptual model of annotated object (details of other parts of the model are left out for clarity)

Building on earlier work (Melgar et al., 2016), in this paper we argue the need for application support for more specificity of the annotation target (see Figure 1). We identify two additional issues with the current state-of-the-art:

4. The W3C annotation protocol lacks support for a potential annotation target identifying and describing itself to the annotation tool.
5. The model also lacks a schema, which would allow scholars or website maintainers to define constraints for a specific class of annotations that is applicable in the context of a specific group of scholarly objects.

Use case: annotation in scholarly digital editions

These issues are illustrated by a scenario of a digital scholarly edition where scholars have a need for annotation support (Boot, 2009, Robinson, 2004, Siemens et al., 2012). Consider an edition that wants to incorporate an external annotation tool into its pages (Figure 2): an edition server shows an edition to a client in a browser. The annotation client runs within that same browser window, but doesn't know about the edition's structure and it talks with its own server. To communicate intelligently with the user, the annotation client needs information about the structure of the edition, which has to be provided by the edition.

The annotation tool should know about the edition's structure for a number of reasons:

- The edition often contains multiple representations of the same text fragment. There might be a diplomatic and a critical transcription, one or more translations, audio versions, and who knows what other versions, and annotations made in one of these should be available in others;
- Other sites may have other editions of this particular text. It should be possible to exchange annotations between them;
- The edition has an internal structure, e.g. a book divided in chapters, or the fragments appearing in modern authors' drafts, or the elaborate structure with multiple apparatuses of some editions of medieval texts. An annotation that refers to a specific component of an edition should be able to address that component and know what sort of component it is.
- The edition should be able to propose suitable annotation types for its components. For personal names, it might suggest an annotation type that links the person to an authority file. For transcriptions, there might be special annotation types for proposed corrections to the transcription. Edition collaboratories could use the annotation functionality to solicit multiple sorts of specialised information from its collaborators.

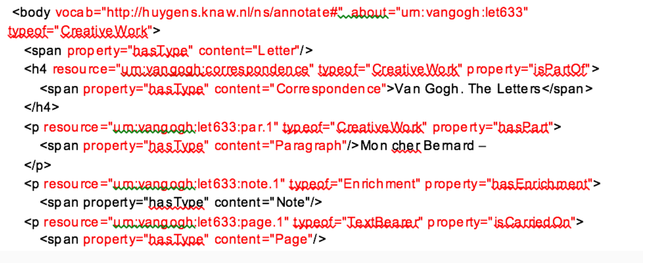
This proposal requires that: (i) the edition describes itself and its structure to the annotation tool, and provides suitable labels for the annotatable objects; (ii) the edition can suggest annotation types for the annotatable objects;

(iii) the effort to integrate annotation functionality in existing editions is minimal; (iv) the annotation tool is generic, but able to handle the created annotations with awareness of the structure that they apply to (it can e.g. return aggregated annotations); (v) the annotation targets are durable and not formulated in terms of HTML structure; and (vi) URI's should be treated as opaque (i.e., we shouldn't try to guess the relations between the annotated components based on their URIs); and lastly (vii) URIs should be canonical.

Proposed Solution

We propose a solution similar to Schema.org (an initiative for adding structural semantics to information on the web) whereby descriptive information about annotatable resources is made accessible to the client by embedding it in the HTML presentation layer through RDFa attributes (Figure 3), using an extensible resource descriptive ontology. Figure 4 shows a basic ontology for text objects (left half of Figure 4) with an edition-specific extension for the example edition (right half of Figure 4). This ontology shares concepts with the FRBRoo ontology (Bekiari et al., 2015) but starts from specific annotation-related concepts. In future work we will investigate extending the ontology with FRBRoo concepts.

Although this approach is focused on annotation of resources on the web, the same principle could be applied in offline annotation, if the offline resources are described in a similar way and annotation clients are developed to make use of this. Also, descriptive information for textual sources can be embedded as markup, but for audiovisual documents, this has to be done via a separate representation, for instance using SMIL (Bulterman et al., 2008).



```
<body vocab="http://huygens.knaw.nl/ns/annotate#" about="urn:vangogh:let633"
typeof="CreativeWork">
   <span property="hasType" content="Letter"/>
   <h4 resource="urn:vangogh:correspondence" typeof="CreativeWork" property="isPartOf">
      <span property="hasType" content="Correspondence">Van Gogh. The Letters</span>
   </h4>
   <p resource="urn:vangogh:let633:par.1" typeof="CreativeWork" property="hasPart">
      <span property="hasType" content="Paragraph"/>Mon cher Bernard –
   </p>
   <p resource="urn:vangogh:let633:note.1" typeof="Enrichment" property="hasEnrichment">
      <span property="hasType" content="Note"/>
   <p resource="urn:vangogh:let633:page.1" typeof="TextBearer" property="isCarriedOn">
      <span property="hasType" content="Page"/>
```

Figure 3. HTML fragments of a letter of Vincent van Gogh (http://vangoghletters.org/orig/let633) described by embedded RDFa. The letter is identified by a URN (urn:vangogh:let633) and defined as a CreativeWork. It is part of a larger CreativeWork, Van Go

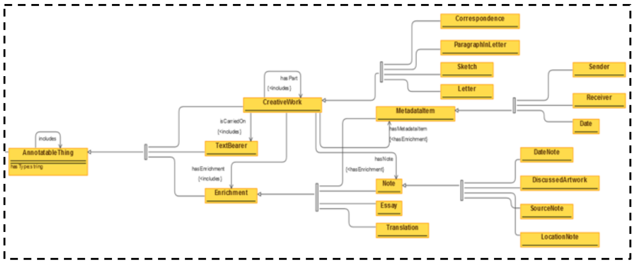


Figure 4. Basic ontology for text objects (left of dashed line) and extended ontology for Van Gogh Letters Collection (right of dashed line). The basic ontology recognizes three types of annotatable things: the creative work being edited and its parts (also creative works), the text bearers (e.g. manuscript pages), and editorial enrichment of any sort. Projects can create an extended ontology to suit their needs. The extended ontology shown here creates specialized classes for the needs of the Van Gogh letter edition (http://vangoghletters.org/).

Methodological impact

In our proposal annotatable resources describe their own semantic structure, thereby facilitating fine-grained annotations. With the RDFa attributes, annotation clients can identify the annotation target in terms of the resource structure (issue 4), which makes annotations less dependent on specific views on the underlying object. Furthermore, this allows development of lightweight open source annotation clients that web services can easily embed to bring annotation to collections of scholarly interest (issue 3).

This makes it easier for scholars to use and reuse annotations to support the argument made in a scholarly article (issue 1). It allows distinguishing different groups of annotations, so researchers can choose to display certain groups of annotations, thereby avoiding being drowned by irrelevant annotations (issue 5). It facilitates employing annotation functionality to ask for targeted comments on resource parts (what do you think of this translation? What clarification of this material are you missing?). Scholars can also more meaningfully combine and compare them across collections and media types, e.g. analyse the correspondence between book and film versions of an intellectual work (issue 2).

If the annotations are consistently stored using open protocols, it becomes possible to reference them in scholarly publications. Collateral benefit of floating this form of 'deep web' semantics to the surface is that other external services such as search engines can also use the exposed semantic information to reason about available resources.

Bibliography

Bekiari, C., Doerr, M., Riva, P., Le Bœuf P. (2015). FRBR, object-oriented definition and mapping from FRBRer, FRAD and FRSAD - International Working Group on FRBR and CIDOC CRM Harmonisation, Version 2.4, November 2015.

Boot, P., (2009). Mesotext: digitised emblems, modelled annotations and humanities scholarship. Amsterdam University Press, 2009.

Bulterman, D., Hansen, J., Cesar, P. et al. (2008). Synchronized Multimedia Integration Language. W3C Recommendation 01 December 2008. https://www.w3.org/TR/2008/REC-SMIL3-20081201/

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., Ledda, G. (2012). "Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries." SDA. 2012.

Hunter, J., Cole, T., Sanderson, R., van de Sompel, H. (2010). The Open Annotation Collaboration: A Data Model to Support Sharing and Interoperability of Scholarly Annotations. Presented at the Digital Humanities 2010. Retrieved from http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-860.pdf

Melgar, L., Blom, J., Baaren, E., Koolen, M., Ordelman, R. (2016). A conceptual model for the annotation of audiovisual heritage in a media studies context. Presented at the AudioVisual Material in Digital Humanities 2016 workshop, Krakow, Poland. Retrieved from https://avindhsig.wordpress.com/workshop-2016-krakow/accepted-abstracts/liliana-melgar-jaap-blom-eva-baaren-marijn-koolen-roeland-ordelman/

Melgar, L., Koolen, M., Huurdeman, H.C., Blom, J. (2017). A Process model of Scholarly Media Annotation. In Proceedings of the 2017 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017

Perkel, J. M. (2015) "Annotating the scholarly web." Nature 528.7580 (2015): 153-154.

Rizkallah, É., (2016). QDA software compatibility: Towards an exchange format with developers for their users. Presented at the Reflecting on the future of QDA software, Rotterdam, The Netherlands.

Robinson, P., (2004). "Where we are with electronic scholarly editions, and where we want to be." Jahrbuch für Computerphilologie Online 4 (2004): 123-143.

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., Sloetjes, H. (2008). An Exchange Format for Multimodal Annotations. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech.

Siemens, Ray, Timney, M., Leitch, C, Koolen, C., Garnett, A. (2012). "Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media." Literary and Linguistic Computing 27.4 (2012): 445-461.

Sloetjes, H. (2014). ELAN: Multimedia Annotation Application. In The Oxford Handbook of Corpus Phonology. Retrieved from http://www.oxfordhandbooks.com.proxy.uba.uva.nl:2048/view/10.1093/oxfordhb/9780199571932.001.0001/oxfordhb-9780199571932-e-019

Unsworth, J. (2000). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this. In Humanities Computing: formal methods, experimental practice symposium, King's College, London.

W3C, (2017a). Web Annotation Data Model. W3C Recommendation 23 February 2017. https://www.w3.org/TR/annotation-model/

W3C, (2017b). Web Annotation Vocabulary. W3C Recommendation 23 February 2017. https://www.w3.org/TR/annotation-vocab/

Walkowski, N-O., Barker, E.T.E., (2014). Digital Humanists are Motivated Annotators. Presented at the Digital Humanities 2014, Lausanne, Switzerland. Retrieved from http://dharchive.org/paper/DH2014/Paper-296.xml

University–Community Digitization Partnerships: Accessing Trans Collections in LGBT Community Archives

Elspeth Brown
elspeth.brown@utoronto.ca
University of Toronto, Canada

Cait McKinney
cait.mckinney@utoronto.ca
University of Toronto, Canada

This paper explores university and community partnerships to preserve and provide online access to LGBTQ cultural heritage materials, focusing in particular on projects that improve the accessibility and profile of transgender materials in traditional gay and lesbian organizations. The paper reflects on the promises and challenges of pursuing trans-positive, activist digital humanities projects within the context of a historically cis gay and lesbian analogue archive. How have anxieties surrounding digital transformations in the archive become enmeshed in concerns about changing understandings of sexual and gender embodiment within the LGBTQ community? How has the queer archive's shift to the digital engendered not only "privacy anxieties" (Chenier 2015) but also identity anxieties, as a new generation of non-binary and trans volunteers, scholars, and activists reshape the meanings of "LGBTQ" within the community archive?

The paper reports on the first three years of the LGBTQ Oral History Digital Collaboratory, a five-year SSHRC-funded project directed by Elspeth Brown at the University of Toronto. The Collaboratory works in partnership with the Canadian Lesbian and Gay Archives, a volunteer-run, community archives, to design and implement in-house digitization programs. This project has invested significant labour and technology into the acquisition and digitization of transgender materials under-represented in the CLGA's collection, and in historically "Gay and Lesbian" community archives more broadly (Rawson 2009; Morris and Rawson, 2013).

This paper will outline the technical and political concerns behind two Collaboratory/CLGA partnership projects aimed at improving access to trans materials: 1) The development of an audio digitization station for oral history cassette tapes; and 2) the processing and digitization of transsexual artist and activist Mirha-Soleil Ross' personal papers. We argue that the planning and implementation of digitization initiatives is a significant catalyst for broader, mandate-driven shifts within LGBTQ cultural heritage organizations. We argue that this is particularly the case when these partnerships drawn on humanities-based approaches that understand the archive as a worlding technology (Stoler 2010, Eichhorn 2013) caught up within broader conflicts that challenge and divide LGBT spaces. More specifically, the paper outlines how university partners can support the development of digital infrastructures in community-based, volunteer-engaged contexts. We suggest that this labour can be understood as part of a broader politics of digitization that must be approached with care when forming community partnerships. We understand digitization as a wide-ranging process that transcends the conversion of "analog" materials into digital formats: digitization has the potential to shift economies of attention in social movement organizations undergoing transition (McKinney 2015).

Theoretically, the paper engages with queer, trans, and feminist approaches to digital archives (Chenier 2009; McLeod et al, 2014; Matte 2015). In particular, we draw on scholarship that has considered how trans materials have the potential to confound, challenge, and ultimately re-formulate archival systems (Brown 2015; Roberto 2011). Here, digital spaces and technologies provide historical "worldmaking" opportunities often denied to trans histories within "LGBT" spaces (Rawson 2014). Emphasizing the relationship between "trans" and movement (Stryker, Currah, Moore 2008), we consider the moments where digitization meets trans as potentially productive but also acrimonious challenges for LGBT archives primarily concerned, in practice, with GayandLesbian (Noble 2006) cultural memory.

The paper covers two case studies from the Collaboratory/CLGA project, and will explain the technical questions and organizational concerns behind the design and roll-out of both projects in order to support the argument outlined above.

Audio Digitization Station

Begun in year one (2014), this station sought to develop a technologically accessible digitization solution for audio-cassette tapes, most of which contain oral histories produced in the 1980s and 90s. We will outline the accessibility concerns at the heart of this project, including how the audio digitization protocol the project established sought to build organizational comfort with digitization practices by integrating the system with existing database infrastructures and ways of organizing volunteer labour at the archives. This part of the paper will also address how the Collaboratory identified trans audio collections within the archives, and will touch on some of the challenges in locating these materials. This material will be useful for humanities scholars in sexuality and gender studies who are concerned about creating digital preservation plans for audio-based primary source materials they might collect or create.

Mirha-Soleil Ross Papers

Begun in year three (2016), this project is working to process, digitize, and improve access to the personal papers of Mirha-Soleil Ross, one of the most significant figures in transgender activism and cultural production in Canada. Donated to the archives in 2008, this large collection had not yet been processed by archives volunteers and as result, was mostly inaccessible to researchers. We are working in collaboration with Ms. Ross, and several members of the trans community, to organize these materials, determine access restrictions for sensitive content, create an online finding aid, and an online digital collection using the Omeka platform, which will sample from the larger volume of material. This part of the paper asks how digitization might provide an opportunity to repair damaged relationships between LGBT organizations and trans communities. It also considers some of the unique privacy concerns scholars must consider when providing online access to materials that document trans lives.

Bibliography

Brown, E. H. (2015). "Trans/Feminist Oral History." *TSQ* 2(4): 666–72.

Chenier, E. (2009). "Hidden from Historians: Preserving Lesbian Oral History in Canada,"*Archivaria*, no. 68: 247 – 70.

Chenier, E. (2015) "Privacy Anxieties: Ethics vs. Activism in Archiving Lesbian Oral History Online." *Radical History Review* 122 : 129-141.

Eichhorn, K. (2013). *The Archival Turn in Feminism: Outrage in Order.* Philadelphia: Temple University Press.

Matte, N. (2015). "Without a Minority/Identity Framework." *Transgender Studies Quarterly,* November 2015, Volume 2, Issue 4: 595-606.

McKinney, C. (2015). "Body, Sex, Interface: Reckoning with Images at the Lesbian Herstory Archives." *Radical History Review* 122: 115–28.

McLeod, D., Rault, J., and Cowan, T.L. (2014) "Speculative Praxis Towards a Queer Feminist Digital Archive: A Collaborative Research-Creation Project," *Ada: A Journal of Gender, New Media, and Technology* #5 ,http://adanewmedia.org/2014/07/issue5cowanetal/?utm_source=rss&utm_medium=rss&utm_campaign=issue5-cowanetal

Morris III, C.E. and Rawson, K.J. (2013) "Queer Archives/Archival Queers." *In Theorizing Histories of Rhetoric.* Ed. Michelle Ballif. (Carbondale, IL: Southern Illinois UP), 74–89.

Noble, B. (2006) *Songs of the Movement: FtMs Risking Incoherence on a Post-Queer Cultural Landscape* (Toronto: Women's Press), 19.

Rawon, K. J. (2009). "Accessing Transgender // Desiring Queer(er?) Archival Logics." *Archivaria* 68: 123-140.

Noble, B. 2014. "Transgender Worldmaking in Cyberspace: Historical Activism on the Internet." *QED: A Journal of Queer Worldmaking* 1(2): 38–60.

Roberto, K.R. 2011. Inflexible Bodies: Metadata for Trasngender Identities. *Journal of Information Ethics* 20(2): 56–64

Stoler, A. L. (2010). *Along the Archival Grain: Epistemic Anxieties and Colonial Common Sense.* Princeton NJ: Princeton University Press.

Stryker, S., Currah, P., and Moore, L. J. (2008). Introduction: Trans-, Trans, or Transgender? *Women's Studies Quarterly* 36(3–4): 11–22.

Cultural (Re-)formations: Structuring a Linked Data Ontology for Intersectional Identities

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Abigel Lemak
alemak@uoguelph.ca
University of Guelph, Canada

Colin Faulkner
cfaulk01@mail.uoguelph.ca
University of Guelph, Canada

Kim Martin
kmarti20@uoguelph.ca
University of Guelph, Canada

Alliyya Mohammed
University of Guelph, Canada

Jade Penancier
University of Guelph, Canada

Rob Warren
rwarren@math.carleton.ca
Carleton University, Canada

Introduction

Cultural diversity has been an increasing source of debate within the digital humanities community. The concentration within the *Debates in Digital Humanities* series (Gold, 2012; Gold and Klein, 2016) of pieces reflecting the increasing prominence of matters related to race, gender, cultural diversity and difference is but one marker of the extent to which diversity matters. The Orlando Project in feminist literary history incorporated an intersectional understanding of identity categories from the outset (Brown, Clements and Grundy, 2006-2017). Translating Orlando's Extensible Markup Language (XML) data into linked open data (LOD) to make it accessible, interoperable, and amenable to a range of analytical approaches (Simpson and Brown) requires an ontology that will serve both Orlando and the broader research community hosted by the Canadian Writing Research Collaboratory (CWRC). This paper outlines the CWRC ontology design and the challenges of shifting from semi-structured to structured data (Smith, 2016: 273).

Much work on digital diversity expresses skepticism of the ability of systematized knowledge structures to capture the performative, processual, and contingent nature of lived subjectivities. Tara McPherson (2012) stresses that "computers are themselves encoders of culture" and calls for more attention to be paid to the interconnectedness of the structures of code and the management of race socially: "Just as the relational database works by normalizing data—that is, by stripping it of meaningful, idiosyncratic context, creating a system of interchangeable equivalencies—our own scholarly practices tend to exist in relatively hermetically sealed boxes or nodes." Scholars including Lisa Nakamura (2002: 120) and Moya Bailey (2011) see value in "messiness" as a way to push against and redefine the contours of a digital humanities scholarship that remains rooted in predominantly white epistemology.

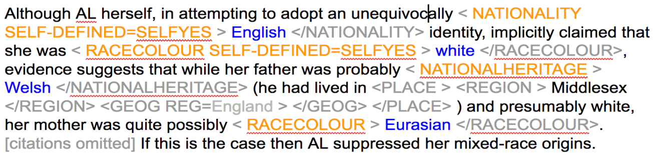
At the same time, relegating representations of difference to narrative rather than structured data will produce gaps within big data that are both impoverishing for humanities inquiry and dangerous in their political implications (Lerman, 2013; Trevinarus, 2014; "Use"; Brown and Simpson, 2013). Adriel Dean-Hall and Robert Warren (2013) have advocated approaches that respect the privacy and preferences of lived human subjects while improving the responsiveness of online systems to diversity and complexity. Within a LOD context, what are finally findable, processable, and reusable on the global graph are things, not strings, so the challenge is the extent to which nuance, context, and indeed messiness can be incorporated into a LOD ontology.

The Orlando Project (Brown, et al., 2006-2017) charted a middle ground between narrative and structure with its bespoke XML tagset. The team struggled with the hierarchical nature of XML, particularly in relation to identity categories, torn between knowledge that readers would turn to Orlando to find writers associated with particular cultural identities and recognition that such categories are discursive rather than essential (Fuss, 2013). We devised a "Cultural Formation" tagset to depict identity as neither unitary nor immutable, and as much related to representational acts as to the lived experiences into which those representations blur. Precisely because constituted through discursive and social practices, vocabularies associated with subjectivities and identities can shift over time and place, and throughout an individual's lifetime.

Cultural formation tagset

The Cultural Formation (CF) tagset recognizes categorization as endemic to social experience, while incorporating variation in terminology and contextualization of identity categories by employing tags at different discursive levels. CF tags describe the subject positions of individuals through 1) contextual tags that encode substantial discussions: class; language; nationality; race and ethnicity; religion; and sexuality; and 2) granular tags that describe, in a word or short phrase, class; ethnicity; gender; geographical heritage; language; nationality; national heritage; political affiliation; race or colour; religious denomination, and sexual identity. With the exception of gender and social class, the Orlando schema eschewed fixed attribute values for the granular tags, allowing the prose to employ the most appropriate language for the context. The structure was not entirely logical or parallel, and we are making the ontology more consistent. The granular tags possess attributes regarding forebears and whether a subject self-identified with a particular term. The tagset aimed to highlight the extent to which social classification is culturally produced and discursively embedded. Rather than disambiguating leaky cultural categories, the team considered them as mutually constitutive with historically specific discursive frameworks, including our tagging structures.

CF encoding pointed users towards a framework for raising and debating complex matters for cultural investigation rather than standardized classifications, refusing to neatly group writers into distinct and fixed categories, since those categories were neither stable nor mutually exclusive (Algee-Hewitt, Porter, Walser, forthcoming). The coding structure can represent quite complex identities, as in the case of Anna Leonowens, the writer whose story of life as governess to the royal Siamese harem was popularized in *The King and I.* Partial markup for the first paragraph of her CF description is shown in Figure 1.



Although AL herself, in attempting to adopt an unequivocally < NATIONALITY SELF-DEFINED=SELFYES > English </NATIONALITY> identity, implicitly claimed that she was < RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR>, evidence suggests that while her father was probably < NATIONALHERITAGE > Welsh </NATIONALHERITAGE> (he had lived in <PLACE > <REGION > Middlesex </REGION> <GEOG REG=England > </GEOG> </PLACE>) and presumably white, her mother was quite possibly < RACECOLOUR > Eurasian </RACECOLOUR>. [citations omitted] If this is the case then AL suppressed her mixed-race origins.

Figure 1: Adapted from Brown, Clements and Grundy, “Anna Leonowens”, Life tab, Show Markup option

The CF component of Orlando’s knowledge representation is thus crucial to its intersectional approach to identity (Brown et al., 2006). Creating a LOD ontology that was not self-referential, however, requires translating the strings or literal values from CF tags, to link Orlando’s semantic structures to other semantic web communities.

LOD ontology creation

An ontology “is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse” (Wikipedia, Ontology - Information Science). Using a standard ontology language such as OWL allows others to interact and exchange with a particular view of the world through a computational process of mediation. As a representation of that understanding, an ontology can be referenced, (dis)agreed with, extended, and used operationally. The coexistence of different representations provides the foundation for translations between LOD concepts.

Ontology creation in our case, as in many others, was driven by the idiosyncrasies and limitations of an existing data set. The information architectures of application databases or XML stores are not always reconcilable to a consistent information system. The CF tagset represents a major challenge in that its structure was designed to eschew disambiguation. Even the major tags were difficult to relate within a concise ontology (Figure 2).

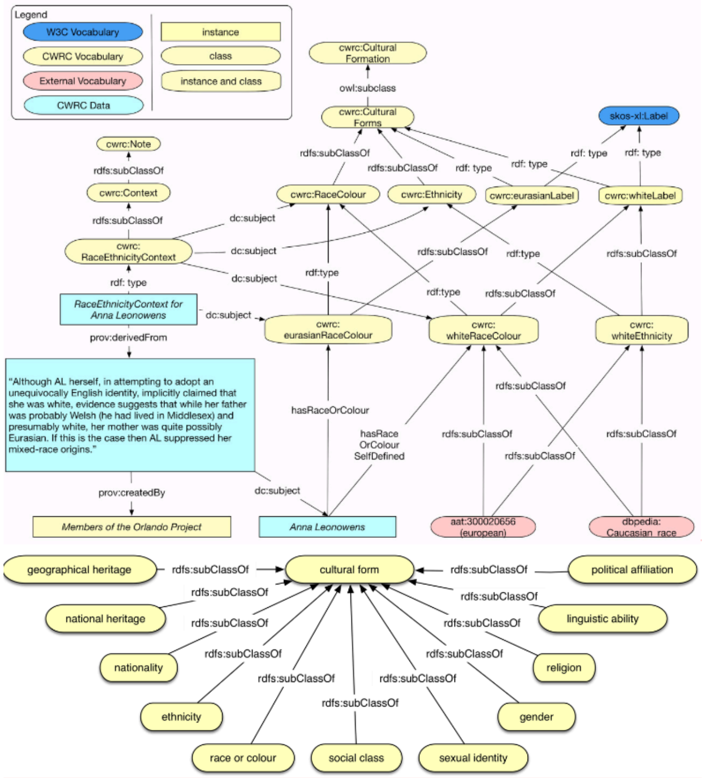


Figure 2: Schematic representation of the granular Cultural Formation tags from Orlando (Please note that these representations are simplified in order to make them legible to the reader.)

For example, nationality and national heritage are not employed as commensurate with citizenship, a well-defined legal concept related to an organized state. They can also be related to a geographical area, which may or may not coincide with a state. Finally, nationhood can reference socio-political constructs such as Lesbian Nation (Johnston, 1973; Ross, 1995; Munt,1998) or disavowals of nationality such as Virginia Woolf’s (1938: 197), which Orlando quotes alongside assigning Woolf an English nationality, a contradiction that requires contextual evidence to make sense.

Linked into context

We decided to make all human-readable annotations within the dataset instances of contextual notes to which the ontological classes are directly tied (Figure 3).

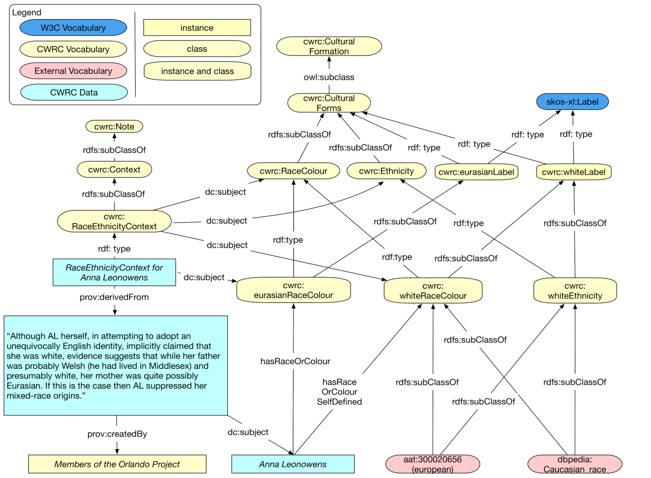


Figure 3: Schematic representation of how the discursive context (note) links to the classificatory structure, and how classificatory labels relate to predicates and external ontologies. Skos:narrower/broader relationships are also used, but omitted here to improve legibility

Thus we model the discursive context within a Race[or]EthnicityContext class. The note instance links to instances of granular category labels, here RaceColour; it provides the provenance and the basis for links to source information. Linking to the provenance of the LOD is particularly important for disputed or contradictory information, as in our example. We are modeling the original Orlando narrative as a source document for our LOD provenance using the the Web Annotation Data Model's subproperty instances. We aim to link every triple to the prose from which it is derived, providing provenance information and contains citations to the sources on which identity assertions are based.

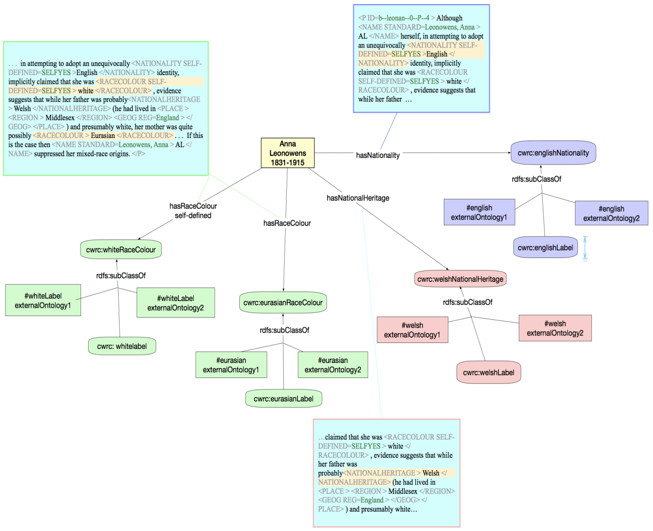
Relating cultural formations

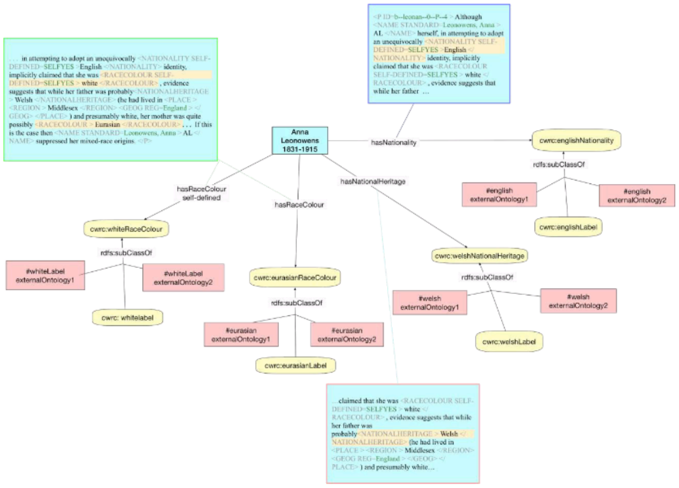
Cultural formation for Orlando is understood primarily as representational, which is not to say that cultural formation is not real or that it has no material effects. The complex signifiers of cultural identities float across Orlando tags as cdata or free text in a semi-structured representation of cultural identities and categories. For the CWRC ontology, we strategized to relate this ontological perspective to that of external vocabularies without conflating our truth with theirs. Our architecture does not import other ontologies wholesale, but adopts components of major vocabularies such as BIBO, FOAF, and FRBR, and relates to large vocabularies in defined ways. As indicated in Figure 3, the instances of cwrc:whiteRaceColour and cwrc:whiteEthnicity within the CWRC ontology are subclasses of the cwrc:whiteLabel. This retains the ambiguity of terms such as “ white” or “ Jewish” precisely as labels that draw together particular types of identity categories, as well as subClasses of those labels. As indicated, those subClasses can be linked to terms in external vocabularies, but both internal and external terms are understood within the CWRC ontology as labels. Indeed, constructing this ontology has brought home to us the need for the LOD community to think through with greater care the relationship between representation and “ reality” in LOD ontologies. A further complication is that identity categories are not only historically contingent but often also change over a particular individual's lifetime. The Orlando dataset supports such nuance in only a few cases, so we have not started with this gnarly problem, but we aim to build into the ontology the capacity to represent such cultural formation dynamics in order to accommodate more temporally precise data.

Conclusion

The CWRC ontology design avoids representing RDF extractions from Orlando data as positivist assertions, and yet produces machine-readable OWL/RDF-compliant graph structures. The ontology allows references to, without endorsing, external ontological vocabularies that are nevertheless part of documenting intersectional cultural processes and identities.

We will present the CWRC ontology as built around the CF design described here, and we will demonstrate its implications through several practical examples. Figure 4 shows schematically the intersectionality of multiple identity categories associated with Leonowens, including the ways that instances are related by subclass relationships in accordance with OWL principles. This importantly allows us to reference components of other ontologies (here the Muninn Appearances ontology, Library of Congress Subject Headings, Getty Art and Architecture Thesaurus, and DBpedia) without adopting them wholesale.


. . . in attempting to adopt an unequivocally <NATIONALITY SELF-DEFINED=SELFYES >English </NATIONALITY> identity, implicitly claimed that she was <RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR> , evidence suggests that while her father was probably<NATIONALHERITAGE > Welsh </NATIONALHERITAGE> (he had lived in <PLACE > <REGION > Middlesex </REGION> <GEOG REG=England > </GEOG> </PLACE>) and presumably white, her mother was quite possibly <RACECOLOUR > Eurasian </RACECOLOUR> . . . If this is the case then <NAME STANDARD=Leonowens, Anna > AL </NAME> suppressed her mixed-race origins. </P>
<P ID=b--leonan--0--P--4 > Although <NAME STANDARD=Leonowens, Anna > AL </NAME> herself, in attempting to adopt an unequivocally <NATIONALITY SELF-DEFINED=SELFYES >English </NATIONALITY> identity, implicitly claimed that she was <RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR> , evidence suggests that while her father …
Anna Leonowens 1831-1915
hasNationality
cwrc:englishNationality
rdfs:subClassOf
#english externalOntology1
#english externalOntology2
cwrc:englishLabel
hasRaceColour self-defined
hasRaceColour
hasNationalHeritage
cwrc:whiteRaceColour
rdfs:subClassOf
#whiteLabel externalOntology1
#whiteLabel externalOntology2
cwrc: whitelabel
cwrc:eurasianRaceColour
rdfs:subClassOf
#eurasian externalOntology1
#eurasian externalOntology2
cwrc:eurasianLabel
cwrc:welshNationalHeritage
rdfs:subClassOf
#welsh externalOntology1
#welsh externalOntology2
cwrc:welshLabel
…claimed that she was <RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR> , evidence suggests that while her father was probably<NATIONALHERITAGE > Welsh </NATIONALHERITAGE> (he had lived in <PLACE > <REGION > Middlesex </REGION> <GEOG REG=England > </GEOG> </PLACE>) and presumably white…



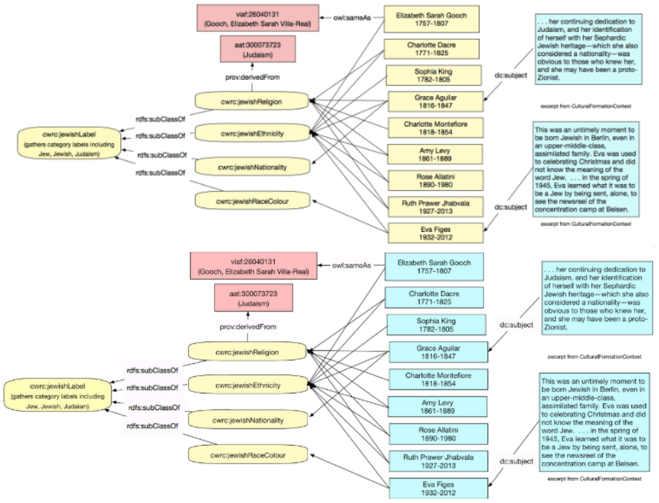
. . . in attempting to adopt an unequivocally <NATIONALITY SELF-DEFINED=SELFYES >English </NATIONALITY> identity, implicitly claimed that she was <RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR> , evidence suggests that while her father was probably<NATIONALHERITAGE > Welsh </NATIONALHERITAGE> (he had lived in <PLACE > <REGION > Middlesex </REGION> <GEOG REG=England > </GEOG> </PLACE>) and presumably white, her mother was quite possibly <RACECOLOUR > Eurasian </RACECOLOUR> . . . If this is the case then <NAME STANDARD=Leonowens, Anna > AL </NAME> suppressed her mixed-race origins. </P>

<P ID=b--leonan--0--P--4 > Although <NAME STANDARD=Leonowens, Anna > AL </NAME> herself, in attempting to adopt an unequivocally <NATIONALITY SELF-DEFINED=SELFYES >English </NATIONALITY> identity, implicitly claimed that she was <RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR> , evidence suggests that while her father ...

...claimed that she was <RACECOLOUR SELF-DEFINED=SELFYES > white </RACECOLOUR> , evidence suggests that while her father was probably<NATIONALHERITAGE > Welsh </NATIONALHERITAGE> (he had lived in <PLACE > <REGION > Middlesex </REGION> <GEOG REG=England > </GEOG> </PLACE>) and presumably white...

Figure 4: Cultural Formation triples related to Anna Leonowens, with corresponding XML-encoded context notes

Figure 5 indicates the ability to see patterns and outliers related to different categorizations of Jewishness in a small subset of Orlando authors.



. . . her continuing dedication to Judaism, and her identification of herself with her Sephardic Jewish heritage—which she also considered a nationality—was obvious to those who knew her, and she may have been a proto-Zionist.

This was an untimely moment to be born Jewish in Berlin, even in an upper-middle-class, assimilated family. Eva was used to celebrating Christmas and did not know the meaning of the word Jew. . . . in the spring of 1945, Eva learned what it was to be a Jew by being sent, alone, to see the newsreel of the concentration camp at Belsen.

Figure 5: Subset of CF triples related to a subset of writers, with sample context annotations and external links; predicates linking individuals to subclasses are inferred (e.g. the edge between Elizabeth Sarah Gooch and cwrc:jewishReligion is hasReligion)

The ontology thus makes all specific instances of Jewishness as religion, ethnicity and so on subclasses of the label for Jewishness that groups them and highlights the extent to which the signifier Jewish is embedded in discourse. Our live presentation will demonstrate the ontology in action using the interactive HuViz (Humanities Visualizer) interface with a larger dataset. This dataset will contain a wider range of identity categories than we have been able to lay out in the diagrams here, and show the interaction between the ontology terms and the representations of individuals.

Ontologies

- CWRC ontology: http://sparql.cwrc.ca/ontology/cwrc
- CWRC sparql end point: http://sparql.cwrc.ca/
- Orlando Biography schema containing Cultural Formation tagset: https://github.com/cwrc/CWRC-Schema/blob/master/schemas/orlando_biography.rng

Bibliography

Alexiev, V., Cobb, J., Garcia, G., and Harpring, P. (2016). *Getty Art and Architecture Thesaurus*. J. Paul Getty Trust. http://vocab.getty.edu/doc/queries

Algee-Hewitt, M., Porter, J. D. and Walser, H. (Forthcoming, 2017). “Representing race and ethnicity in American fiction: 1789-1964.”

Bailey, M.Z. (2011). "All the digital humanists are white, all the nerds are men, but some of us are brave." *Journal of Digital Humanities* 1.1. http://journalofdigitalhumanities.org/1-1/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/ (accessed 7 April 2017)

Brickley, D., and Miller, L. (2000-2014). *FOAF Vocabulary Specification 0.99*. http://xmlns.com/foaf/spec/

Brown, S., Clements, P., and Grundy, I (eds.) (2006-2017). *Orlando: Women’s Writing in the British Isles from the Beginnings to the Present.* Cambridge: Cambridge University Press Online.

Brown, S., Clements, P., and Grundy, I. (2006). "Sorting things in: Feminist knowledge representation and changing modes of scholarly production." *Women's Studies International Forum* 29.3.

Brown, S., & Simpson, J. (2013, October). The curious identity of Michael Field and its implications for humanities research with the semantic web. In *Big Data, 2013 IEEE International Conference on* (pp. 77-85). IEEE.

Canadian Writing Research Collaboratory. (n.d.) http://cwrc.ca

D’Arcus, B., and Giasson, F. (2008-2013). *Bibliographic Ontology Specification (BIBO)*. http://purl.org/ontology/bibo/ Structured Dynamics.

Davis, I., and Newman, R. (2005). *Functional Requirement for Bibliographic Records (FRBR)* http://purl.org/vocab/frbr/core#

DBpedia. (n.d.)http://wiki.dbpedia.org/

Dean-Hall, A. and Warren, R. (2013). “Sex, privacy, and ontologies.” *SEXI*. Rome, Italy. http://www.dbdump.org/~warren/publications/dean-hall:sexi:2013/dean-hall:sexi:2013.pdf (accessed 7 April 2017).

Fuss, D. (2013). *Essentially Speaking: Feminism, Nature & Difference*. New York: Routledge.

Gold, M. (ed.) (2012). *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press.

Gold, M. K., and Klein, L. F. (eds.) (2016). *Debates in the Digital Humanities 2016*. Minnesota: University of Minnesota Press.

Johnston, J. (1973). *Lesbian Nation: The Feminist Solution.* New York: Simon and Schuster.

Lerman, J. (2013). “Big data and its exclusions.” 66 *Stanford Law Review Online* 55: 55-63. http://www.heinonline.org.subzero.lib.uoguelph.ca/HOL/Page?handle=hein.journals/slro66&start_page=55&collection=journals&id=66 (accessed April 7, 2017).

McPherson, T. (2012). "Why are the Digital Humanities so white? Or thinking the histories of race and computation." In M. Gold (ed). *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press, pp. 139-160.

Muninn Project. “Appearances Ontology Specification - 0.1.” 2012. http://rdf.muninn-project.org/ontologies/appearances.html

Munt, S. (1998). "Sisters in exile: the lesbian nation." *New Frontiers of Space, Bodies and Gender.* London: Routledge, pp. 3-19.

Nakamura, L. (2002). *Cybertypes: Race, Ethnicity, and Identity on* *the Internet.* London: Routledge.

Ross, B. (1995). *The House that Jill Built: A Lesbian Nation in Formation*. Toronto: University of Toronto Press.

Smith, J. (2016). “Working with the Semantic Web.” In C. Crompton, R. J. Lane, and R. Siemens (ed.). *Doing Digital Humanities: Practice, training, research.* London: Routledge, pp. 273-88.

Treviranus, J. (2014). “The value of the statistically insignificant.” *Educause* 49:1. http://er.educause.edu/articles/2014/1/the-value-of-the-statistically-insignificant (accessed 7 April 2017).

W3C. (2017). Web Annotation Data Model. 23 February 2017. https://www.w3.org/TR/annotation-model/ (accessed: 7 April 2017).

Wikipedia contributors (2017). "Ontology (information science)," *Wikipedia, The Free Encyclopedia.*https://en.wikipedia.org/w/index.php?title=Ontology_(information_science)&oldid=772391479 (accessed April 7, 2017).

Woolf, V. (1938). *Three Guineas*. London: Hogarth Press.

The Course of Emotion in Three Centuries of German Text— A Methodological Framework

Sven Buechel
sven.buechel@uni-jena.de
JULIE Lab, Friedrich Schiller University Jena, Germany

Johannes Hellrich
johannes.hellrich@uni-jena.de
JULIE Lab, Friedrich Schiller University Jena, Germany

Udo Hahn
udo.hahn@uni-jena.de
JULIE Lab, Friedrich Schiller University Jena, Germany

Introduction

Texts not only carry factual, but also affective information, such as expressions of joy or grief. In the humanities, especially literary studies, emotion expression and elicitation (often in texts from prior language stages) have been focused on in many contributions (see, e.g., Carroll and Gibson (2011), Poppe (2012), Hillebrand (2011)).

Similarly, in natural language processing (NLP), emotion analytics have developed into an active area of research (Liu, 2015). Nevertheless, there is little previous work explicitly addressing emotion in historical language and the specific methodological problems this raises. Hamilton et al. (2016) as well as Cook and Stevenson (2010) presented methods for identifying amelioration and pejoration of words. Acerbi et al. (2013) and Bentley et al. (2014) demonstrated the potential of emotion analysis for the Digital Humanities (DH) by linking temporal emotion patterns in texts to major sociopolitical events and trends in the 20th century.

Our work goes beyond these studies in two ways: we claim to be more adequate as we combine these two approaches to analyze non-contemporary **text** based on time-specific **lexical** resources. We also claim to be more informative as we employ the Valence-Arousal-Dominance (VAD) model of emotion (Bradley and Lang, 1994) instead of simple **polarity** (positiveness/negativeness) alone. We have already shown the latter to be beneficial in DH applications (Buechel et al., 2016a). We hope that our work will be a step towards a new set of tools especially beneficial for areas such as literary studies (e.g., in distant reading (Moretti, 2013)) or history of mind.

Methods

The VAD model of emotion assumes that affective states can be described relative to three emotional **dimensions**, i.e., Valence (corresponding to the concept of polarity, see above), Arousal (the degree of excitement or calmness) and Dominance (feeling in or out of control of a social situation). The VAD dimensions allow for a more fine-grained modeling than polarity alone, e.g., words like *orgasm* and *relaxed* have similar Valence but opposing Arousal values (Bradley and Lang, 1999). Formally, the VAD model constitutes a three-dimensional vector space illustrated by Figure 1 (Buechel and Hahn, 2016).

The association of words with a VAD score is captured in **emotion lexicons**. These can either be empirically determined by aggregating subjective judgments of several human subjects; or they can be semi-automatically constructed allowing for greater size but reducing accuracy on individual words. For the semi-automatic construction, the typical approach is to expand an existing lexicon (the **seed lexicon**) based on word similarity (see below). There are several competing expansion algorithms. Cook and Stevenson (2010) were the first to describe expansion algorithms for the induction of the emotion value of words for non-contemporary language stages by using word similarity values determined from historical corpora.

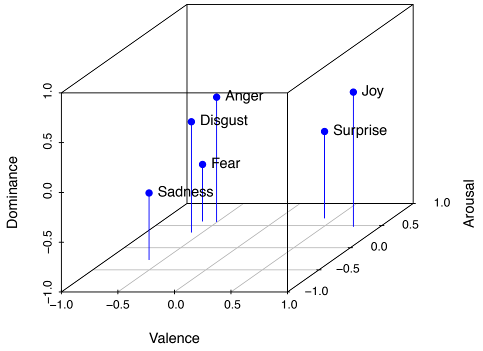


Figure 1: The VAD vector space. For ease of understanding, the positions of six *Basic Emotions* (Ekman, 1992) are given.

Extending this approach, we compared several emotion induction algorithms, viz., those by Turney and Littman

(2003), Hamilton et al. (2016), and Bestgen (2008). The former two were slightly modified to make them deal with numerical input values (for a more detailed description of these methods, see Buechel et al. (2016b) and Hamilton et al. (2016)).

We used point-wise mutual information with singular value decomposition (Levy et al, 2015; SVD$_{\mathrm{PPMI}}$) to measure word similarity, since it turned out to be superior for DH applications in previous work (Hellrich and Hahn, 2016). We used the German ANGST lexicon (Schmidtke et al., 2014) as seed. The individual algorithms were evaluated by comparing our induced historical lexicons against judgments of knowledgeable PhD students from the humanities. For this task, we asked them to make their assessments **as if** they were contemporary readers from the respective time period. The Turney-Littman algorithm performed best in this set-up and was thus employed for all subsequent analyses.

Experiments

For demonstration purposes, we here apply our methodology to the core corpus of the *Deutsches Textarchiv* (DTA; Geyken, 2013, TCF version from May 11, 2016) [German Text Archive], a well-curated and balanced collection of historical German texts. We analyzed texts created between 1690 and 1899, splitting the resulting corpus into seven slices (each spanning 30 years) to achieve similarly sized and sufficiently large subcorpora for further processing. We computed word similarities within each of these slices and then applied the Turney-Littman expansion algorithm, thus creating seven distinct emotion lexicons, each describing the emotion of words for its specific period. Given these temporally stratified lexicons, we claim that shifts in emotion association of words can be traced over time by comparing the emotion values a word exhibits in different lexicons. To validate this claim, we selected the words for which we could compute similarity scores in each time step (as these methods are more accurate for high-frequency words, rare words were excluded from our study) and standardized their VAD values for each lexicon and dimension (VAD).

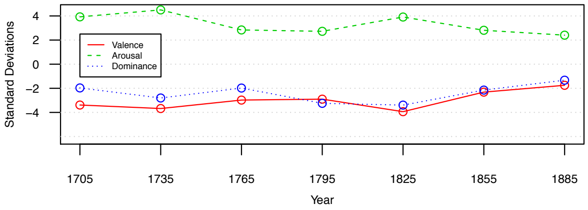


Figure 2: Development of the lexical item *Sünde* [sin] during the study period relative to the VAD dimensions.

Rank	Lemma and Translation			
	1690s		1890s	
1	todt-	(German prefix for 'death' as in 'deadly sins')	Lamm	lamb
2	Erzürnung	infuriation	hinwegnehmen	to take away
3	läßlich	minor (clerical)	Verzeihung	forgiveness
4	beichten	to confess	Ausschweifung	excess
5	Nachlaß	abatement/ inheritance	Gotte	god
6	Grobheit	crudeness	Schande	shame
7	verschweigen	to conceal	Reue	repentance
8	beweinen	to beweep	Ärgernis	nuisance
9	pichen	to pitch	Laster	vice
10	beichten	to confess	aufrichtig	sincere

Table 1: Top ten collocations of the lexical item *Sünde* [sin] in the DTA corpus comparing the 1690s and the 1890s using pointwise mutual information for scoring. Source: http://kaskade.dwds.de/dstar/dta/diacollo/

We illustrate this approach with an example from Figure 2. It displays an overall amelioration of *Sünde* [sin] whose onset roughly coincides with the age of enlightenment—often understood as the starting point of secularization (Sheehan, 2003), although care must be taken when interpreting these word course graphs since noise can be introduced from various sources (such as word similarity and emotion induction algorithms); both strongly depend on the amount of data available for each time step. This observation is in line with well-known findings from descriptive lexicography (Grimm and Grimm, 1942). The same semantic shift can also be discovered by a more established method, namely collocation analysis.

Table 1 reveals that *Sünde,* at the end of our examination period, has acquired an additional moral-bourgeois meaning facet (implied, e.g., by *Ausschweifung* [excess], *Ärgernis* [nuisance] and *Laster* [vice]) which was not present in the beginning. There, only the religious sense is traceable.

Going one step further, we then used these lexicons to examine how emotion is distributed over literary texts in the DTA in the course of time. We employed the *Jena Emotion Analysis System* (JEMAS; Buechel and Hahn, 2016), an open-source tool for emotion measurement using a configurable VAD lexicon. We processed each DTA text with the period-aligned lexicon, leading to the main methodological contribution of our work: linking the research areas of automatically inducing historical **word** emotion (e.g., Hamilton et al., 2016) and emotion prediction in historical **text** (e.g., Acerbi et al., 2013).

We scaled the resulting emotion values within each VAD dimension tracing the development of the three principal literary forms—Narrative, Lyric, and Drama—in German literature between 1690 and 1899. For each of the seven 30-year periods (organized in rows), we created three scatterplots (one for each pair of the VAD dimensions; organized in columns) resulting in 21 plots in total (Figure 3). Each data point represents one text—color and shape represent membership to the respective form.

It is evident from the plots how the distinction of the individual forms increases and decreases in emotional terms in the course of time. This application differs from typical stylometric approaches since we employ emotional features instead of word counts. We find the most distinct emotional patterns between 1780 to 1809 (approximately covering the Weimar Classicism) and between 1870 to 1899 (covering the late German Realism). Drama displays consistently more Arousal than Lyric and Narrative since 1750, whereas Lyric seems to be the most positive class (Valence) expressing the least control (Dominance). Of course, the examination of the DTA offers many more intriguing findings, however, for brevity, we limit ourselves here to presenting examples.

Conclusion

In this contribution, we described a novel methodological framework for quantifying emotion in non-contemporary text. Applying this approach to a 210-years section of the German DTA corpus, we find clear emotional signals for temporally evolving distinctions between the principal literary forms, viz. Narrative, Lyric, and Drama. All resources and software we developed for this work are publicly available.

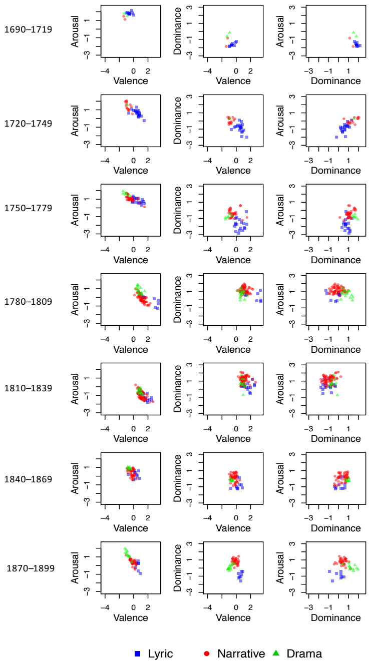


Figure 3: Distribution and development of the principal literary forms, Lyric (blue), Drama (green) and Narrative (red), relative to each pair of VAD emotions (in columns) between 1690 and 1899 (each row representing a 30-year slice).

Acknowledgments

This research was partially conducted within the Graduate School "The Romantic Model" supported by grant GRK 2041/1 from the Deutsche Forschungsgemeinschaft (DFG).

Bibliography

Acerbi, A., Lampos, V., Garnett, P. and Bentley, R.A. (2013). The expression of emotions in 20th century books. *PLoS ONE* 8(3): e59030.

Bentley, R.A., Acerbi, A., Ormerod, P. and Lampos, V. (2014). Books average previous decade of economic misery. *PLoS ONE*, 9(1): e83147.

Bradley, M.M. and Lang, P.J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1): 49–59.

Bradley, M.M. and Lang, P.J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective Ratings. Report C–1, The Center for Research in Psychophysiology, University of Florida, Gainesville.

Bestgen, Y. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pp. 496–500.

Buechel, S. and Hahn, U. (2016). Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. *Proceedings of the 22nd European Conference on Artificial Intelligence*, pp. 1114–22.

Buechel, S., Hahn, U., Goldenstein, J., Händschke, S.G.M. and Walgenbach, P. (2016a). Do enterprises have emotions? *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 147–53.

Buechel, S. Hellrich, J. and Hahn, U. (2016b). Feelings from the past—Adapting affective lexicons for historical emotion analysis. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities*, pp. 54–61.

Carroll, N., and Gibson, J. (eds) (2011). *Narrative, Emotion, and Insight.* University Park: Pennsylvania State University Press.

Cook, P. and Stevenson, S. (2010). Automatically identifying changes in the semantic orientation of words. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pp. 28–34.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4): 169–200.

Geyken, A. (2013). Wege zu einem historischen Referenzkorpus des Deutschen: das Pro- jekt Deutsches Textarchiv. *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pp. 221– 34.

Grimm, J. and Grimm, W. (eds) (1942). Sünde. In: *Deutsches Wörterbuch*, volume 20.

Hamilton, W.L., Clark K., Leskovec, J. and Jurafsky D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 595–605.

Hellrich, J. and Hahn, U. (2016). Bad company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: Technical Papers, pp. 2785–96.

Hillebrand, C. (2011). *Das emotionale Wirkungspotenzial von Erzähltexten: Mit Fallstudien zu Kafka, Perutz und Werfel*. Berlin: Akademie Verlag.

Levy, O., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3: 211–25.

Liu, B. (2015). *Sentiment Analysis: Mining Opinion, Sentiments and Emotions*. New York: Cambridge University Press.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Poppe, S. (2012). *Emotionen in Literatur und Film*. Würzburg: Königshausen & Neumann.

Schmidtke, D.S., Schröder, T., Jacobs, A.M. and Conrad, M.

(2014). ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods*, 46(4): 1108–18.

Sheehan, J. (2003). Enlightenment, religion, and the enigma of secularization: A review essay. *The American Historical Review*, *108*(4): 1061–80.

Turney, P.D. and Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4): 315–46.

A Ten-Year Summary of a SOA-based Micro-services Infrastructure for Linguistic Services

Marco Büchler
mbuechler@etrap.eu
University of Goettingen, Germany

Greta Franzini
gfranzini@etrap.eu
University of Goettingen, Germany

Emily Franzini
efranzini@etrap.eu
University of Goettingen, Germany

Thomas Eckart
teckart@informatik.uni-leipzig.de
Universität Leipzig, Germany

Introduction

In the mid 1990s, the Natural Language Processing Group at the University of Leipzig began work on the Wortschatz project which aims to provide corpora in hundreds of languages and in different size-normalisations, be that 100K, 300K or 1M sentences. As the resources grew in size, so did the number of requests for the data. In the early stages of the project a specific dump was created, parts of which even came with a small user-interface. The database dump was shared with interested researchers and partners in the business sector.

After some time, however, the personnel costs of this kind of collaboration became unsustainable. For this reason, a new plan was put into motion in 2004, consisting of the development of a SOAP-based API - the *Leipzig Linguistic Services* (LLS) - that enabled any interested person to access the data of the *Wortschatz* databases in any provided language (Quasthoff et al. 2006, Eckart et al. 2012). Overall 20 services were provided, delivering specific information such as baseform, category classifications, and thesaurus data. The aim of the LLS was to establish a *Service Oriented Architecture* (SOA) for linguistic resources based on small and atomic micro-services that could be combined by users for particular needs. Users were then not only able to browse through the *Wortschatz* website, but also to integrate those services with their own existing digital ecosystems.

In 2005 these services were made publicly available and by September 2006 all requests were systematically logged. In July 2014 the number of logged requests reached nearly one billion. While at the beginning the use was limited to academia, over time the services were increasingly used by the private and business sectors as well.

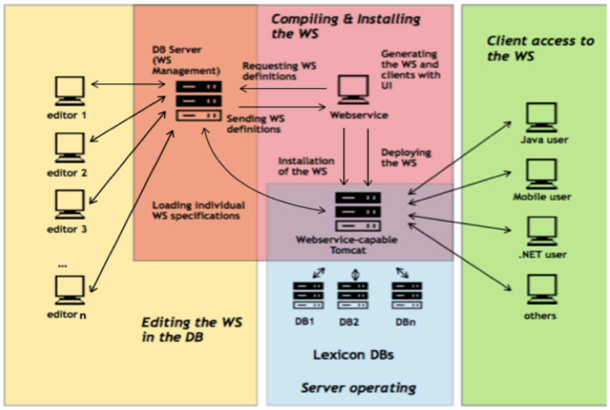


Figure 2. Four workflow modes separated by concern: editing (yellow); managing, compiling and deploying (red); hosting and operating (blue); using the LLS infrastructure (green).

The Leipzig Linguistic Services

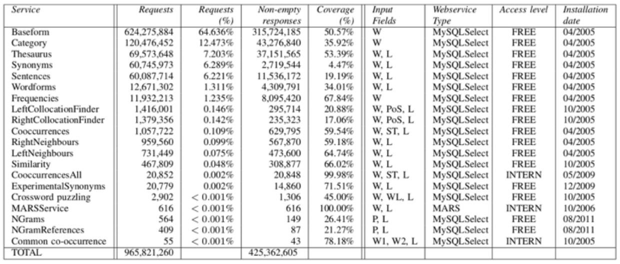
The intention of the overall LLS architecture was to be as simple and generic as possible. A generic architecture can be reused in different scenarios but tends to have too many parameters and options, while a simple architecture claims usability and guarantees a faster learning curve. In the following, we briefly describe the architecture of the LLS.

In order to create the server-side Java code for a specific webservice, a data-set needed to be added to the webservice management (yellow zone in figure 1). The necessary edits contain, besides others, information on the name and type of the webservice (see also table 1) or parameters. *Apache Ant* was used as the central tool for generating the back-end services and deploying them in a *Tomcat* server (see red zone in figure 1). The blue zone illustrates the operations of the *Wortschatz* databases. Using the generic description of the webservice in the WSDL-files a number of wrappers of generated source code were created and made publicly available by LLS users such as for C# as part of .NET, Perl, Python, Delphi, PHP, Ruby and JavaScript (see green zone in figure 1).

Independently from the underlying programming languages, over the past ten years we have observed different uses in research, business and in the private sector. In research, the LLS were used in the areas of text profiles and author classification (Borchardt 2005). The services were also used as data resources for sentiment analysis or for query expansion. Users from the business field were mainly interested in using *Baseform* or *Synonym* services for improving internal search indexes. The LLS data was also used for information retrieval tasks in portals for weighting words in a word cloud or to display enriching information. Private users accessed the LLS to complete crossword puzzles. A dedicated service was installed upon request just for this purpose (see also table 1), since it was possible to query a pattern of an incomplete word with a given word length limitation. From 2008 the SOA-based cyberinfrastructure of LLS was re-used in Digital Humanities projects such as eAQUA and eTRACES (Büchler et al. 2008).

Results

Table 1 provides an overview of the 20 services offered with a breakdown of the requests and the responses. Over half of the requests (*64.6%*) were made to the *Baseform* service. Similarly, services with high-quality and often manually-curated data, such as the *Thesaurus* and *Synonyms* services, were requested more often than the quantitatively-computed *Similarity* service, which provided similarly used words by assuming the distributional hypothesis (Harris 1954), and thus compared the co-occurrence vectors of two words. Even if the coverage for this service, *66.02%*, is significantly higher than, for example, the *Category* (*35.92%*) or the *Synonyms* (*4.47%*) services, users appeared to prefer precision over recall for their end-user applications.



Service	*Requests*	*Requests (%)*	*Non-empty responses*	*Coverage (%)*	*Input Fields*	*Webservice Type*	*Access level*	*Installation date*
Baseform	624,275,884	64.636%	315,724,185	50.57%	W	MySQLSelect	FREE	04/2005
Category	120,476,452	12.473%	43,276,840	35.92%	W	MySQLSelect	FREE	04/2005
Thesaurus	69,573,648	7.203%	37,151,565	53.39%	W, L	MySQLSelect	FREE	04/2005
Synonyms	60,745,973	6.289%	2,719,544	4.47%	W, L	MySQLSelect	FREE	04/2005
Sentences	60,087,714	6.221%	11,536,172	19.19%	W, L	MySQLSelect	FREE	04/2005
Wordforms	12,671,302	1.311%	4,309,791	34.01%	W, L	MySQLSelect	FREE	04/2005
Frequencies	11,932,213	1.235%	8,095,420	67.84%	W	MySQLSelect	FREE	04/2005
LeftCollocationFinder	1,416,001	0.146%	295,714	20.88%	W, PoS, L	MySQLSelect	FREE	10/2005
RightCollocationFinder	1,379,356	0.142%	235,323	17.06%	W, PoS, L	MySQLSelect	FREE	10/2005
Cooccurrences	1,057,722	0.109%	629,795	59.54%	W, ST, L	MySQLSelect	FREE	04/2005
RightNeighbours	959,560	0.099%	567,870	59.18%	W, L	MySQLSelect	FREE	04/2005
LeftNeighbours	731,449	0.075%	473,600	64.74%	W, L	MySQLSelect	FREE	04/2005
Similarity	467,809	0.048%	308,877	66.02%	W, L	MySQLSelect	FREE	10/2005
CooccurrencesAll	20,852	0.002%	20,848	99.98%	W, ST, L	MySQLSelect	INTERN	05/2009
ExperimentalSynonyms	20,779	0.002%	14,860	71.51%	W, L	MySQLSelect	FREE	12/2009
Crossword puzzling	2,902	< 0.001%	1,306	45.00%	W, WL, L	MySQLSelect	FREE	10/2005
MARSService	616	< 0.001%	616	100.00%	W, L	MARS	INTERN	10/2006
NGrams	564	< 0.001%	149	26.41%	P, L	MySQLSelect	FREE	08/2011
NGramReferences	409	< 0.001%	87	21.27%	P, L	MySQLSelect	FREE	08/2011
Common co-occurrence	55	< 0.001%	43	78.18%	W1, W2, L	MySQLSelect	INTERN	10/2005
TOTAL	965,821,260		425,362,605					

Table 1. Overview of requests made to LLS between 2006-2014, in descending order. The Responses columns only list responses whose value was not empty. For space constraints, the values in the Input Fields column are abbreviated: Word (W.), Limit (L.), Pa

Low coverage is also caused by requests to German language databases, especially by compound nouns that cannot all be included in a *Baseform* or *Category* service. Many multi-word units (MWU) were also requested. Out of all the requests, *84,760,875* (*8.78%*) were MWUs. With regard to the distribution of the webservice usage, only the two most frequently requested services, *Baseform* and *Category*, were queried more often than the total count of the MWU requests. This speaks to the impact of MWUs.

The less frequently used webservices in table 1 were primarily limited to internal uses, to newly installed services or, as was the case for the Crossword Puzzling service, to manual usage instead of automatic bulk requests.
The following questions are discussed in the paper:

1. Geographical distribution and spread of requests
2. Requested languages distribution
3. Requests by cleanliness in terms of broken encodings or sending HTML code
4. Temporal distribution including lessons learnt from incompatibility issues of used software and their new versions causing a decrease in service usage
5. Identified service chains of the atomic LLS micro-services that users built on the client-side
6. Experiences for load balancing of linguistic services
7. Interoperability issues of programming languages and interpreting the WSDL-files differently
8. Comparisons of SOAP- and REST-based webservices

Conclusion

“If you build it, they will come“ is an infrastructure mantra that we can answer given the atomic micro-services of the LLS (more critical view by van Zundert 2012). However, with regard to easy-to-integrate and atomic micro-services we found that users were generally very pragmatic as they requested everything that they had found in texts or on webpages, such as RGB colour-sets, URLs and other meta-information. Based on the log-files, we conclude that it is easier to request a token and look for a match in the LLS database of millions of words rather than to invest only little time in conventional pre-processing and pre-selection on the client-side. Similarly, users repeatedly requested function words, sometimes only a few minutes apart. This user behaviour entailed a significant server load and user control over the requests. This type of recurring request on unchanged data could only be considered as spam.

We found that providing an infrastructure like the LLS over the course of a decade challenges the compatibility of used software components.

Moreover, from a Natural Language Processing (NLP) standpoint, the results contribute to existing conversations about the difficulty of building balanced and representative corpora. In fact, user interests detected in the LLS log-files can help to enrich corpora by adding further topics. The contribution also touches upon discussions about qualitative and manually-curated data versus automatically-computed and quantitatively-available results of language technology algorithms. Notwithstanding the improvement of NLP algorithms, our results show that users prefer qualitative data and that they often request these services even if the domain and concept coverage is relatively low. The conclusion we draw from the user behaviour observed in almost one billion requests is that research fields, including the Digital Humanities, should share their data –no matter how small– through large infrastructure initiatives like DARIAH and CLARIN in order to increase the textual coverage of linguistic resources.

Bibliography

Borchardt, S. (2005) *Generierbarkeit einer XML Topic Map aus E-Mails unter Verwendung von Text-Mining-Methoden und Nutzung von Web Services*. Bachelor thesis.

Büchler, M., Heyer, G., Gründer, S. (2008) *Bringing Modern Text Mining Approaches to Two Thousand Years Old Ancient Texts e-Humanities* At: Workshop in the 4th IEEE International Conference on e-Science.

Eckart, T., Quasthoff, U., and Goldhahn, D. (2012) *Language Statistics-Based Quality Assurance for Large Corpora*, Proceedings of Asia Pacific Corpus Linguistics Conference.

Harris, Z. (1954) Distributional structure, Word, 10, 2-3, pp. 146162.

Quasthoff, U., Richter, M., and Biemann, C. (2006) *Corpus Portal for Search in Monolingual Corpora* Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC).

Van Zundert, J. (2012) , *If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities*, Historical Social Research / Historische Sozialforschung, vol. 37, no. 3, pp. 165-86.

Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg

Daniel Schlör
daniel.schloer@informatik.uni-wuerzburg.de
University of Würzburg

Ulrike Henny
ulrike.henny@uni-wuerzburg.de
University of Würzburg

Christof Schöch
christof.schoech@uni-wuerzburg.de
University of Würzburg

Summary

We propose a way to work with the stylometric distance measure Delta to analyse the subgenre of texts written by different authors. For that, we neutralize the author signal by penalizing the texts from the same writer, allowing the texts to have their shortest distances to other authors' works. We test this method with several subcorpora of Spanish prose and a corpus of French theatre.

Stylometry and Delta beyond Authorship

Since John Burrows proposed it in 2002, Delta has been one of the most used and researched methods in stylometry and authorship attribution. Burrows explained it as "expression of difference, pure difference" (2002: 269) and is based on basic statistical concepts like most frequent words, z-scores and the Manhattan distance between each pair of texts. Burrows closes his paper with an unanswered question about why Delta works so well.

Other researchers such as Hoover (2004b: 454), Argamon (2008), Plasek (2014) or Evert et al (2015: 79) have confirmed that we are still far from being able to answer this question. This lack of understanding has not stopped the stylometric community of trying to improve Delta (Hoover 2004a; Argamon 2008; Eder 2013). Smith and Aldridge (2011) have proposed Cosine Delta which gives the best results in different languages (Jannidis et al. 2015).

Since Delta is sensitive to aspects or *signals* like genre or period (Burrows 2002), the corpora for authorship attribution tend to be homogenous in those aspects. Research has been conducted to try to separate signals (Schöch 2013 and 2014) or selecting the words that contribute to them using Recursive Feature Elimination (Büttner and Proisl 2016). Jannidis and Lauer (2014) and Hoover (2014) show how Delta can be used to distinguish genre and periods within the works of a single author. Other researchers have used other methods such as classification (Hettinger et al., 2016; Underwood 2014) or logistic regression (Jockers 2013; Riddell and Schöch 2014) to similar ends.

Neutralizing Author Signal in Delta

Our proposal is to neutralize the author signal directly on the Delta matrix. We use a testing corpus of texts from three Spanish authors and three subgenres. Detailed information about the corpora, files, parameters and scripts is in our GitHub repository. We applied Cosine Delta (5000 MFW) with Stylo (Eder, Rybicki and Kestemont 2016) and visualized the resulting distance matrix with Python:

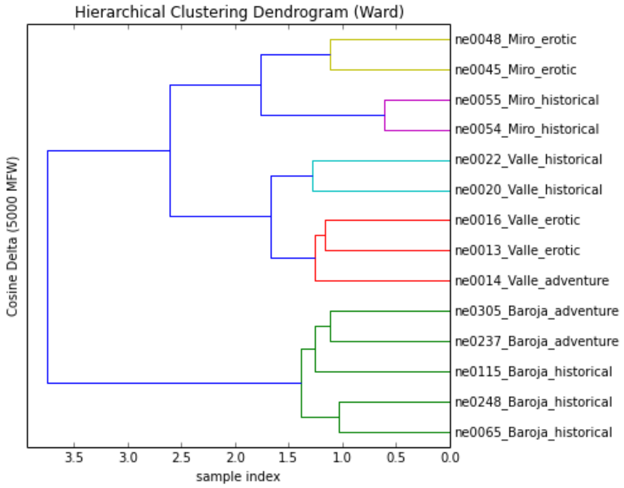
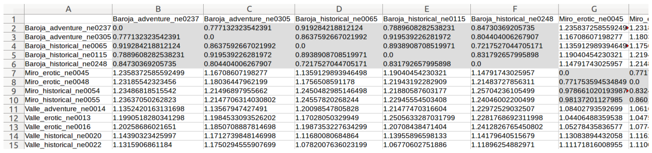


Figure 1: Dendrogram from Cosine Delta

As expected, the texts are clustered by author, with subclusters of subgenres. The underlying Delta Matrix contains distances between all texts:



	A	B	C	D	E	F	G	
1		Baroja_adventure_ne0237	Baroja_adventure_ne0305	Baroja_historical_ne0065	Baroja_historical_ne0115	Baroja_historical_ne0248	Miro_erotic_ne0045	Miro_
2	Baroja_adventure_ne0237	0.0	0.777132323542391	0.919284218812124	0.7889608282538231	0.84730369205735	1.235837258559249	1.231
3	Baroja_adventure_ne0305	0.777132323542391	0.0	0.8637592667021992	0.919539226281972	0.804404006267907	1.16708607198277	1.180
4	Baroja_historical_ne0065	0.919284218812124	0.8637592667021992	0.0	0.8938908708519971	0.7217527044705171	1.135912989394649	1.175
5	Baroja_historical_ne0115	0.7889608282538231	0.919539226281972	0.8938908708519971	0.0	0.831792657995898	1.19040454230321	1.219
6	Baroja_historical_ne0248	0.84730369205735	0.804404006267907	0.7217527044705171	0.831792657995898	0.0	1.14791743025957	1.214
7	Miro_erotic_ne0045	1.2358372585592499	1.16708607198277	1.1359129893946498	1.19040454230321	1.14791743025957	0.0	0.771
8	Miro_erotic_ne0048	1.23185542323456	1.18036447962199	1.1756508591178	1.21943192282909	1.21483727856311	0.771753594534849	0.0
9	Miro_historical_ne0054	1.23486818515542	1.21496897955662	1.2450482985146498	1.21880587603177	1.25704236105499	0.978661020193987	0.832
10	Miro_historical_ne0055	1.23637050262823	1.2147706314030802	1.24557820268244	1.22945554503408	1.24046002200499	0.98137201127985	0.860
11	Valle_adventure_ne0014	1.1352420163131698	1.13567947427491	1.20098547805828	1.21477470316604	1.22972529032507	1.08402793592699	1.061
12	Valle_erotic_ne0013	1.1990518280341298	1.1984533093526202	1.17028050329949	1.2505633287031799	1.2281768692311998	1.04406488359538	1.047
13	Valle_erotic_ne0016	1.20258686021651	1.1850708887814698	1.1987353227634299	1.20708438471404	1.2412826765450802	1.05278435836577	1.077
14	Valle_historical_ne0020	1.14390323425997	1.1712739848146998	1.11680080684864	1.13955896598133	1.14179640515679	1.13083894432058	1.116
15	Valle_historical_ne0022	1.1315906861184	1.1750294555907699	1.0782007636023199	1.06770602751886	1.11896254882971	1.11171816008955	1.110

Figure 2: Cosine Delta Matrix

We see a tendency of lower Delta values for documents of the same author (below 1.0) in comparison to documents of different authors (above 1.0). But what about the closest texts written by a different author? For the historical novel in column E, they are in the rows 14 and 15 and are historical novels, as well. This pattern is found for the majority of the texts. How could we cluster the texts preferring the closest text from other authors? And if we are able to neutralize the author signal, will we see noise or subgenre clusters?

Our proposal is to penalize the distances between the texts of the same author (cf. Lu and Leen 2007 for penalization in image clustering), making them closer to the average distance of texts of different authors, then cluster the neutralized distance matrix and measure the cluster homogeneity by author and subgenre.

We define the set of all documents by an author *a* as A_a, the collection containing all documents by all authors as *C* and total number of documents in the collection is defined as *c*:

$$A_a := \{d_1, \cdots, d_{m_a}\}$$

$$C = \{A_1, \cdots, A_n\}$$

$$c := |\bigcup C|$$

Note that each document is in exactly one author-document set A_i.

First, we calculate the average distance of texts of all pairwise different authors (in fig. 2, all the distances in black). We call this value the **mean of different authors or *M(C)*** and for this collection its value is 1.16.

$$M(C) := \frac{\sum\limits_{\substack{A_a, A_b \in C, a \neq b \\ d_i \in A_a, d_j \in A_b}} \Delta(d_i, d_j)}{\sum\limits_{A} |A| \cdot (c - |A|)}$$

Second, we calculate the **mean of the texts of each author a *M(A_a)*** (in fig. 2, the distances in grey).

$$M(A_a) := \frac{\sum\limits_{\substack{d_i, d_j \in A_a \\ i \neq j}} \Delta(d_i, d_j)}{|A_a| \cdot |A_a - 1|}$$

For each author, we subtract his/her mean value from the mean of different authors *M(C)* - *M(A_a)* resulting in the **difference of the author**. This value represents how far the texts of a specific author are to the mean of different authors:

author	mean	difference
Miro	0.607	0.552
Baroja	0.669	0.490
Valle	0.752	0.407

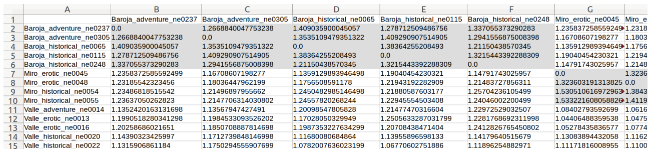
Figure 3: Means and differences of author

Third, we add the difference of the author *M(C)* - *M(A_a)* to the Delta values of text of the same author. This gives a Neutralized Delta-function as follows:

$$\forall d_i \in A_a, d_j \in A_b$$

$$\tilde{\Delta}(d_i, d_j) := \begin{cases} \Delta(d_i, d_j) & \text{for } a \neq b \\ \Delta(d_i, d_j) + (M(C) - M(A_a)) & \text{for } a = b \end{cases}$$

This converts the table from Figure 1 into a Neutralized Delta matrix:



	A	B	C	D	E	F	G	
1		Baroja_adventure_ne0237	Baroja_adventure_ne0305	Baroja_historical_ne0065	Baroja_historical_ne0115	Baroja_historical_ne0248	Miro_erotic_ne0045	Miro_e
2	Baroja_adventure_ne0237	0.0	1.2668840047753238	1.409035900045057	1.278712509486756	1.337055373290283	1.235837258559249	1.2318
3	Baroja_adventure_ne0305	1.2668840047753238	0.0	1.3535109479351322	1.409290907514905	1.2941556875008398	1.16708607198277	1.1803
4	Baroja_historical_ne0065	1.409035900045057	1.3535109479351322	0.0	1.38364255208493	1.21150438570345	1.135912989394649	1.1756
5	Baroja_historical_ne0115	1.278712509486756	1.409290907514905	1.38364255208493	0.0	1.3215443392288309	1.19040454230321	1.2194
6	Baroja_historical_ne0248	1.337055373290283	1.2941556875008398	1.21150438570345	1.3215443392288309	0.0	1.14791743025957	1.2148
7	Miro_erotic_ne0045	1.2358372585592499	1.16708607198277	1.1359129893946498	1.19040454230321	1.14791743025957	0.0	1.3236
8	Miro_erotic_ne0048	1.23185542323456	1.18036447962199	1.1756508591178	1.21943192282909	1.21483727856311	1.323603191313825	0.0
9	Miro_historical_ne0054	1.23486818515542	1.21496897955662	1.2450482985146498	1.21880587603177	1.25704236105499	1.530510616972963	1.3843
10	Miro_historical_ne0055	1.23637050262823	1.2147706314030802	1.24557820268244	1.22945554503408	1.24046002200499	1.533221608058826	1.4119
11	Valle_adventure_ne0014	1.1352420163131698	1.13567947427491	1.20098547805828	1.21477470316604	1.22972529032507	1.08402793592699	1.0616
12	Valle_erotic_ne0013	1.1990518280341298	1.1984533093526202	1.17028050329949	1.2505633287031799	1.2281768692311998	1.04406488359538	1.0475
13	Valle_erotic_ne0016	1.20258686021651	1.1850708887814698	1.1987353227634299	1.20708438471404	1.2412826765450802	1.05278435836577	1.0774
14	Valle_historical_ne0020	1.14390323425997	1.1712739848146998	1.11680080684864	1.13955896598133	1.14179640515679	1.13083894432058	1.1162
15	Valle_historical_ne0022	1.1315906861184	1.1750294555907699	1.0782007636023199	1.06770602751886	1.11896254882971	1.11171816008955	1.1100

Figure 4: Author-Neutralized Delta matrix

The values in grey are now in general above 1.0: the texts of the same author have been separated, showing relations between texts independently of authorship. Now the adventure and historical novels of Baroja in columns C and D have their closest text in works of different authors but belonging to the same subgenre.

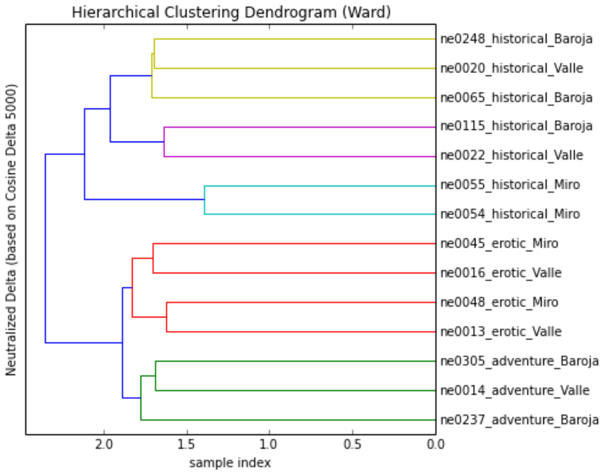


Fig. 5: Author-neutralized Delta dendrogram

In comparison with Figure 1, this dendrogram allows us to see new text relations beyond authorship but within subgenre, showing clusters with different authors but the same subgenre: for example, the cluster of historical novels by Baroja and Valle or the two very close subclusters of erotic novels by Miró and Valle.

Tests and Evaluation

For the evaluation, the homogeneity of the clusters (Rosenberg and Hirschberg, 2007) was measured. This measure yields values between 0 and 1. As ground truth, the metadata about author and subgenre have been used. The results for the dummy corpus:

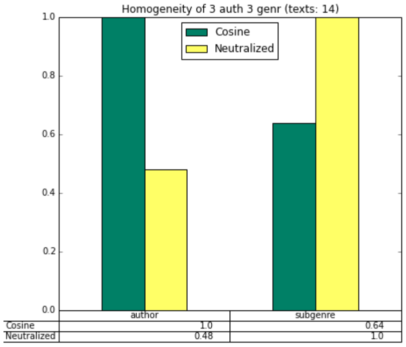


Figure 6: Homogeneity of Cosine and Neutralized Delta for author and subgenre

The homogeneity of the clusters of Cosine Delta (see fig. 1) are perfect for authors and much lower for subgenre, because the author clusters contain subgenre subclusters. The homogeneity of the clusters of Neutralized Delta (see fig. 5) is lower for authorship (as expected), but not for subgenre. In this case the neutralization of the author signal only deteriorates the homogeneity for authorship but improves the homogeneity for subgenre.

We have analysed different subgenres present in the whole corpus for test the method. We created subcorpora of historical, bildungsroman, erotic and adventure novels:

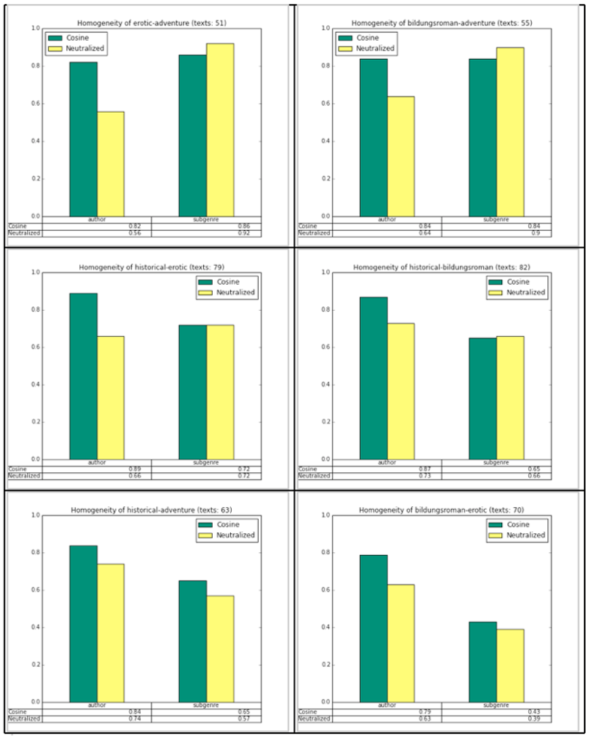


Figure 7: Homogeneities for Spanish prose subcorpora

As expected, the neutralization consistently deteriorates the homogeneity for author (between -0.26 and -0.1) while the homogeneity for subgenre is not deteriorated (between -0.08 and 0.06). The homogeneity for subgenre of adventure compared to erotic and bildungsroman get the best results (over 0.9) and they even improved on results with Cosine. Adventure novels are also best recognized in classification tasks (Hettinger et al. 2016). Subgenres which are very difficult to differentiate like historical and adventure (Pedraza Jiménez and Rodríguez Cáceres 1983: 672 and 1987: 459) get one of the worst results.

The results are similar when testing other corpora, such as a corpus of French drama (Schöch et al. 2015) and a corpus of Spanish American novels:

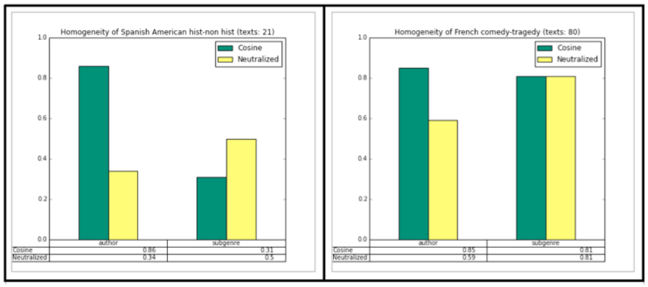


Figure 8: Homogeneity values for French drama and Spanish American novels

Conclusion and future work

Our main goal was to present a method to neutralize the Delta distances of the same author using the difference between the mean of the author and the mean of different authors. Tested on eight subcorpora, this procedure, as we expected, deteriorates the homogeneity of authorship clusters but maintains the subgenre homogeneity, improving it for some cases. That discovers relations between texts (see fig. 5) that were hidden by authorship. This procedure brings a new way of working with Delta beyond authorship attribution.

Both Cosine and Neutralized Delta show very different results for the comparison of different subgenres, something which points to the different internal structure of the subgenres. The comparison of very different subgenres (like adventure against erotic or bildungsroman) gets higher subgenre cluster homogeneity. Neutralized Delta could be used for comparing different corpora of specific subgenres and test the significance of the results to better characterize these subgenres. In an ideal scenario, we would like to test on a perfect balanced corpus where a set of authors are represented in all subgenres of the same period.

For future work, we will analyse how different parameters like versions of Delta or number of MFW affect the results. We also plan to transfer the approach to an earlier step in the Delta procedure and penalize the word z-score vectors.

We look forward to the feedback of the international DH community about this new use of the very effective “expression of difference, pure difference” which is Delta.

Acknowledgements

To avoid confusion regarding intellectual property, we would like to make it clear that the main idea and implementation are the work of the first author. Other authors have brought important remarks, feedback, some of the corpora and have helped with the redaction and the formalisations.

Bibliography

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, **23**(2): 131–47.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, **17**(3): 267–87 http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta.

Büttner, A. and Proisl, T. (2016). Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular. *DHd 2016: Modellierung, Vernetzung, Visualisierung*. Leipzig: Universität Leipzig, pp. 66–69 http://www.dhd2016.de/abstracts/sektionen-002.html.

Calvo Tello, J. (2016). Entendiendo Delta desde las Humanidades. *Caracteres. Estudios culturales y críticos de la esfera digital*, **5**(1): 140–76.

Eder, M. (2013). Bootstrapping Delta: a safety-net in open-set authorship attribution. *DH2013*. Lincoln: UNL https://sites.google.com/site/computationalstylistics/preprints/m-eder_bootstrapping_delta.pdf?attredirects=0.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. *The R Journal*, **16**(1): 1–15 https://journal.r-project.org/archive/accepted/eder-rybicki-kestemont.pdf.

Evert, S., Proisl, T., Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Towards a better understanding of Burrows's Delta in literary authorship attribution. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver CO: Association for Computational Linguistics, pp. 79–88 .

Hettinger, L., Reger, I., Jannidis, F. and Hotho, A. (2016). Classification of Literary Subgenres. *DHd 2016*. Leipzig: Universität Leipzig, pp. 154–58 http://dhd2016.de/boa.pdf.

Hoover, D. L. (2004a). Testing Burrows's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Hoover, D. L. (2004b). Delta Prime?. *Literary and Linguistic Computing*, **19**(4): 477–95.

Hoover, D. L. (2014). A Conversation Among Himselves: Change and the Styles of Henry James. In Hoover, D. L., Culpeper, J. and O'Halloran, K. (eds), *Digital Literary Studies*. New York & London: Routledge, pp. 90–119.

Jannidis, F. and Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In Erlin, M. and Tatlock, L. (eds), *Distant Readings. Topologies of German Culture in the Long Nineteenth Century*. Rochester: Camden House, pp. 29–54 gerhard-lauer.de/index.php/download_file/view/335/1/

Jannidis, F., Pielström, S., Schöch, C. and Vitt, T. (2015). Improving Burrows' Delta – An empirical evaluation of text distance measures. *DH 2015*. Sydney: ADHO http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html.

Jockers, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

Lu, Z. and Leen, T. K. (2007). Penalized Probabilistic Clustering. *Neural Computation*, **19**(6): 1528–67

Pedraza Jiménez, F. B. and Rodríguez Cáceres, M. (1983). *Manual de literatura española. 7: Época del realismo*. Pamplona: Cénlit.

Pedraza Jiménez, F. B. and Rodríguez Cáceres, M. (1987). *Manual de Literatura Española. 9: Generación de Fin de Siglo: Prosistas*. Pamplona: Cénlit.

Plasek, A. (2014). Incommensurability? Authorship, Style, and the Need for Theory. DH2014: Lausanne: ADHO http://dharchive.org/paper/DH2014/Paper-755.xml.

Riddell, A. and Schöch, C. (2014). Progress through Regression. *Digital Humanities DH2014:*. Lausanne: ADHO http://dharchive.org/paper/DH2014/Paper-60.xml.

Rosenberg, A. and Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. Prague: Association for Computational Linguistics, pp. 410–20 https://aclweb.org/anthology/D/D07/D07-1043.pdf.

Schöch, C. (2013). Fine-tuning Our Stylometric Tools: Investigating Authorship and Genre in French Classical Theater. *DH2013*. Lincoln: UNL http://dh2013.unl.edu/abstracts/ab-270.html.

Schöch, C. (2014). Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In Schöch, C. and Schneider, L. (eds), *Literaturwissenschaft im digitalen Medienwandel*. pp. 130–57 http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf.

Schöch, C., Henny, U., Calvo Tello, J. and Popp, S. (2015). *The CLiGS Textbox*. Würzburg: University of Würzburg https://github.com/cligs/textbox.

Smith, P. W. H. and Aldridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. *Journal of Quantitative Linguistics*, **18**(1): 63–88

Underwood, T. (2014). Understanding Genre in a Collection of a Million Volumes, Interim Report. https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251

Mechanizing the Humanities? King's Digital Lab as Critical Experiment

Paul Caton
paul.caton@kcl.ac.uk
King's College London, United Kingdom

Ginestra Ferraro
ginestra.ferraro@kcl.ac.uk.
King's College London, United Kingdom

Luis Figueira
luis.figueira@kcl.ac.uk
King's College London, United Kingdom

Elliott Hall
elliott.hall@kcl.ac.uk
King's College London, United Kingdom

Neil Jakeman
neil.jakeman@kcl.ac.uk
King's College London, United Kingdom

Pam Mellen
pamela.mellen@kcl.ac.uk
King's College London, United Kingdom

Anna-Maria Sichani
anna-maria.sichani@huygens.knaw.*nl*
Huygens ING, Netherlands

James Smithies
james.smithies@kcl.ac.uk
King's College London, United Kingdom

Miguel Vieira
jose.m.*vieira@kcl*.ac.uk
King's College London, United Kingdom

Tim Watts
tim.j.watts@kcl.ac.uk
King's College London, United Kingdom

Carina Westling
c.e.i.westling@sussex.ac.uk
King's College London, United Kingdom

Introduction

The digital humanities (DH) are perhaps unique amongst humanities endeavours. They force us to confront the conceptual and ethical implications that attend the union of the humanities with engineering and organizational thinking. They demand attention to tools, methods, ethics, and pedagogy, but also organizational bureaucracy, human resource management, economics, and systems maintenance. Rather than merely prompting a bland mechanization of the humanities, as critics suggest, this offers a fascinating epistemological challenge. It challenges us to rethink how human meaning and knowledge are constructed, and how they will be remade as the twenty-first century progresses. This requires a step-change in our epistemic, ethical, and collective assumptions as much as our methods.

These issues are distilled in laboratory settings. Sociologist of science Karin Knorr-Cetina (1999) notes that labs function as blended communities, uniting researchers with technicians and administrators. Access to equipment, chemicals, data, and funding, impact the production of knowledge as much as pure research questions. Issues of finance and organizational power compete with creativity and the need for diversity. Scientists have been gaining insight into this since the nineteenth century, fostering traditions (and injustices) that are well understood . (Adas, 1999). Digital humanities labs challenge us to create parallel traditions, appropriate to the humanities and GLAM communities. King's Digital Lab, launched in November 2016, embraces this challenge, positioning itself as an experiment in infrastructural as well as conceptual terms.

Postphenomenological perspectives current in the philosophy of technology and Science and Technology Studies (STS) help explain our approach. Writers like Donald Ihde (2009b) and Peter Kroes (Verbeek, 2010) embrace the entanglement of humans with technological tools, systems, and processes, and meditate on the material reality that informs our experience of the world. They reject the Heideggerean critique of technology, based on the modernist division of subject and object, in favour of acceptance that entanglement with culture, technology, and ideology is not only unavoidable but provides a window into the nature of human experience (Ihde 2009a). Rather than mechanizing the soul of the humanities, digital humanities laboratories force us to confront our entanglement with technology, along with its enabling infrastructures and ideologies.

King's Digital Lab

King's Digital Lab (KDL) builds on a 30-year legacy in digital humanities at King's College London. The lab represents one half of a new digital humanities model, in conjunction with the Department of Digital Humanities (DDH). KDL provides software development and infrastructure to departments in the faculties of Arts & Humanities and Social Science and Public Policy, focusing on software engineering and implementing the systems and processes needed to produce high quality digital scholarly outputs. The department focuses on delivering quality teaching to its postgraduate students and growing cohort of undergraduates, and producing research outputs in line with its status as an academic department. In combination KDL and DDH include close to 40 staff, host 160 projects, served 130 million webpage views in 2014, and teach over 200 students.

KDL's business, operational, and human resource plans define its research values alongside its business and technological model. It has been established with 12 permanent full-time staff: director; project manager; analysts, software engineer, developers, designers; and systems manager. Contract and temporary staff are used as infrequently as possible, ideally to offer student experience in a software development environment. The HR model is explicitly designed to foster sustainable #alt-ac Research Software Engineering (RSE) career paths. All KDL team members, permanent or contract, are encouraged to use 10% of their time on personal projects (either on their own or in collaboration with colleagues), leading to work with Raspberry PIs, virtual reality, and an interest in maker culture.

The KDL model is based on deeply felt humanistic values, but reflects a level of organization required to manage entanglement with technological systems. The lab manages over 90 projects, including up to 20 that are active in some form, and ~5 million digital objects. The team manage over

180 virtual machines, on an infrastructure that uses 400GB of RAM and 27TB of data. New infrastructure platforms are being trialed that include access to cloud and high performance computing options, in a nod towards a future working with big data, visualization, and simulation. The goal is to facilitate a transition from twentieth to twenty-first century modes of computationally intensive humanities and social science research, but to do so in consciously humanistic terms.

In a rejection of a simplistically 'mechanized' future, development tools are proactively managed and the lab has a 'design first' philosophy (Verbeek, 2006). This is partly a way to manage the considerable complexities that come with advanced DH research and the delivery and management of multiple projects, but it also recognizes that digital tools and methods are, at their best, beautiful. Aesthetic and quality values can extend from front-end design to technological platforms, code, tools, and methods. Data, similarly, can and should be beautiful, not only in adherence to appropriate technical standards but in its conformance to scholarly best practice and deep domain knowledge. Infrastructure and systems, likewise, are always compromised by their material design and ideology (Russell, 2014), but decisions to choose open source components and emphasize access and sustainability enhances control and agency (Friedman et al, 2015).

To reduce complexity and improve sustainability, the lab uses the Python programming language, and Django web framework in preference to other options. The full technology stack is consciously oriented towards open source components, and a balance between functionality and sustainability.

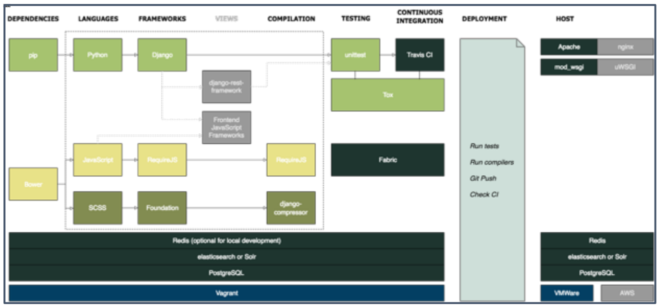


Figure 1

This level of organization helps us manage technology, but also promotes critical awareness: the current technological state of the lab is far from perfect, but it is under control and guided by known critical values. The concept of the 'laboratory' is important in this context. Rather than mechanization, it implies experimentation and risk, but also a certain intellectual seriousness. Scientists learned what a laboratory means to them over a century ago; the humanities and social sciences are only just starting to explore the implications. They are profound, not only in terms of the epistemological implications of putting tools between the researcher and the object of study (with inevitable technical constraints), but in terms of the ideological implications of using industry approaches to software development and financial management. Consciousness of this ensures the lab is sustainable, and can continue to support scholarship as well as the careers of our team members.

The technological inheritance of the lab is considerable. It includes over 90 live web-based projects, built using heterogeneous tools and programming languages by the (historic) Centre for Computing in the Humanities, (historic) Centre for eResearch in the Humanities, and the Department of Digital Humanities. Funding agencies paid for them to be built, but not to sustain them. Some Primary Investigators (PIs) have retired, or are no longer in contact with King's. Support for these projects is currently borne by the lab, generously supported by the Faculty of Arts & Humanities, but is being managed by an evolving archiving and sustainability plan that will assess each of the projects, determine their intellectual merit, and work with their owners to find the best way to maintain or archive them. The archiving and sustainability model used for this task will be published, as well as being included in the lab's Software Development Life-cycle (SDLC), to ensure sustainability will be considered on day one of every new project.

The organizational chart of King's Digital Lab is flat rather than hierarchical, reflecting an aspiration to be role-based and collaborative: a shared intellectual and scholarly space that exists to experiment with new approaches as well as deliver projects on time and budget. The scale is such that the lab design has needed to be out-sourced to multiple authors: director and project manager working with line management to define the business plan and financial model, analysts and developers developing the software life-cycle, UI developer leading the design vision, systems manager ensuring the infrastructure and networking model is appropriate. Together, it amounts to something complex and technologically dependent, but redeemed through a philosophy of shared ownership and conscious experimentation.

Bibliography

Adas, M. (1989). *Machines as the Measure of Men: Science, Technology, and Ideologies of Western Dominance*. Cornell Studies in Comparative History. Ithaca: Cornell University Press

Friedman, B., Kahn, P.H., and Borning, A. (2015). "Value Sensitive Design and Information Systems." In *Human-Computer Interaction and Management Information Systems: Foundations*. New York: Routledge

Ihde, D. (2009) 'Foreword', in Jan Kyrre
Berg Olsen et al, *New Waves in Philosophy of Technology*. Basingstoke: Palgrave Macmillan, 2009), p.xii

Ihde, D. (2009). *Postphenomenology and Technoscience: The Peking University Lectures*. SUNY Series in the Philosophy of the Social Sciences. Albany: SUNY Press.

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, Mass: Harvard University Press.

Russell, A.L. (2014). *Open Standards and the Digital Age: History, Ideology, and Networks*. Cambridge: Cambridge University

Press.
Verbeek, P. (2006). "Materializing Morality Design Ethics and Technological Mediation." *Science, Technology & Human Values* 31 (3): 361–80.
Verbeek, P. (2010). *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Penn State Press.

Networks of the Great Awakenings: Classification of Puritan Sermons by Word Usage Statistics

Michał Choinski
michalchoinski@gmail.com
Jagiellonian University, Poland

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Poland

Revivalism has always played a significant role in the social functioning of religion in the US. As argued by McClymond, "religious revivals are as American as baseball, blues music, and the stars and stripes" (2010, 306). This strong presence of revivalism in the American religious landscape translates into the considerable significance of revival preaching not only for the American pulpit practice, but also for American culture in general. One could argue that a certain continuity is discernible in the American revival tradition, and preachers of consecutive "Awakenings", starting with the Great Awakening, have utilized similar communicative strategies, analogous sets of cultural references, as well as persuasive ploys to forward the "New Birth" to their hearers and to spread the revival zeal. Billy Graham, the most celebrated televangelist of the 20th century, testified to the importance of this tradition when in 1949 in Los Angeles, during the "Canvas Cathedral" Crusade, he delivered to a contemporary audience the most notorious sermon of the Great Awakening (and, perhaps, of America's entire pulpit oratory), Jonathan Edwards's *Sinners in the Hands of an Angry God*. Graham was by no means a pioneer in this respect, as the 19th century American revivalists, like Charles Finney, eagerly fell back on the rhetorical heritage of the first Great Awakening sermons.

Crowds of thousands of people, the emotional reactions of the audiences which often bordered on mass hysteria, fervent theological debates and a surplus of publications played a significant role in the shaping of the early American ecclesiastical order. Similarly, the American rhetorical tradition was strongly informed by the revival developments in pulpit oratory, especially in the context of both the oratory of the American Revolution and the Civil War. New forms of preaching manifested the power of the spoken word, and propelled the dynamics of public debate in a period which was to prove vital for the development of American identity.

The study of American preaching tradition from the diachronic perspective seems particularly important. Different groups of preachers in consecutive Great Awakenings, (First 1735-1750, Second 1790-1840, Third 1850-1900, Fourth 1960-1980) appropriated the rhetorical models employed by the previous generations and built upon their output.

However, because of the sheer size of the available corpus, the comparative study of the preachers of all Great Awakening has been so far impossible. Stylometry based on word usage makes it possible to highlight the connections between particular groups of preachers, as well as to demonstrate the evolution of the American revival preaching tradition; and to confront these findings with the existing attempts at classification/chronology (Stout 1986).

In our study, 42 collections of sermons by individual preachers (one text file per preacher) have been collected, digitized and modernized to avoid an excessive chronological bias due to spelling differences; of course, purely linguistic bias could not be entirely eliminated. Of these, 13 are traditionally classified as First Awakening; 9 as Second; 19 as Third; the most traditionally controversial group, Fourth, is represented by Billy Graham alone.

Classification was made by comparing distances, or differences, between most-frequent-word frequencies of the texts using Burrows's Delta procedure (2002); the distance measure applied was "Cosine Delta" as proposed by Jannidis et al. (2015), implemented in the *stylo* package (Eder et al. 2015, 2016) for R (). The word frequencies were submitted to a consensus procedure of cluster analysis at word frequency vectors of 100 to 2000 most frequent words, and these results then served to produce network diagrams in *Gephi* (Bastian et al. 2009) using the gravitational Force Atlas 2 algorithm (Jacomy et al. 2014). This produces a network or a "map" of data points for the individual preachers' sermons; the closer and the thicker the links between them, the more similar they are.

Figure 1 presents such a network graph for the 42 preachers. The color coding follows the traditional division into four "Awakenings" (from First, green, through Second, yellow, Third, red, and Fourth, purple). A very strong evolutionary pattern emerges; apart from minor imperfections, the network shows a clear evolution from the earliest preachers to the most modern one, Graham. At the same time, the grouping of the data points suggest that a different classification might also exist in the dataset. This is why another algorithm available in *Gephi*, "Modularity," was used to discover these "communities," or groups (Blondel et al. 2008).

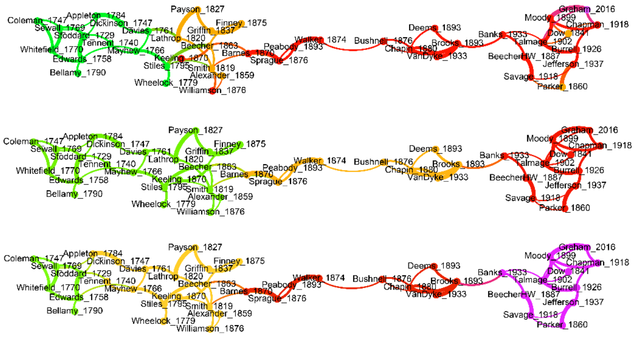


Figure 1. The four Great Awakenings in traditional classification (top); divided by modularity into 3 (center) and 4 (bottom) groups.

This produces two alternatives to the traditional division of the Revivalist movements(s). In the first of these (center), what is usually referred to as the First and the Second Great Awakening would be merged into a single group, while the traditional Third Awakening splits into early and late phases; the latter of which now also included the single representative of the Fourth. Perhaps more interestingly, computer-generated four Awakening communities (bottom) suggest a First only limited to some of Edwards's contemporaries, a Second that extends a little more than traditionally into the past, and again a limited early Third with an expanded Fourth.

This is of course not to say that the above results invalidate the traditional, or historical, division of American revivalist writing. But the stylistic (or, at least, stylometric) divisions are no less valid. Most-frequent-word usage is a significant element of distant reading; and such a distant reading of the Revivalism opens new avenues for close reading of the American homiletic tradition.

Acknowledgements

This research is part of project K/PBO/000331, supported by Poland's National Science Center.

Bibliography

Bastian, M., Heymann, S., Jacomy, M. (2009). *Gephi: an open source software for exploring and manipulating networks.* International AAAI Conference on Weblogs and Social Media.

Blondel, V., Guillaume J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (10), P10008 (12pp) doi: 10.1088/1742-5468/2008/10/P10008.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17: 267-287.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013: Conference abstracts*, University of Nebraska-Lincoln, 487-489.

Eder, M., Rybicki, J., Kestemont, M. (2016). "Stylometry with R: A Package for Computational Text Analysis," *R Journal* 8 (1): 107-121.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, 9(6): e98679. doi:10.1371/journal.pone.0098679.

McClymond, M. (2010). Revivals, in P. Goff (ed.), *The Blackwell Companion to Religion in America*, Chichester: Wiley-Blackwell.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien, http://www.R-project.org/

Stout, S. (1986). *The New England Soul. Preaching and Religious Culture in Colonial New England*. New York: Oxford University Press.

Remembrance of Contemporary Events: On Setting Up The Sunflower Movement Archive

Tyng-Ruey Chuang
trc@iis.sinica.edu.tw
Academia Sinica, Taiwan

Abstract

In the late evening of March 18, 2014, students and activists stormed into and occupied the main chamber of Taiwan's Legislature. The event set off the Sunflower Movement, signifying a turning point in Taiwan's history. Researchers at Academia Sinica arranged to acquire all the supporting artifacts and documentary materials in the chamber before the protest came to a peaceful end. In this paper, we discuss the issues in archiving and making available to the public a large collection of artifacts created by thousands of participants in a contemporary event. We demonstrate systems designed to encourage people to identify items of their own in the archive. We show how an accessible catalog to the archive can help people tell their stories hence collectively may strengthen the public's recollections about the movement.

Background

In the late evening of March 18, 2014, a small group of students and young activists stormed into the main chamber of the Taiwan's Legislature in protest of the hastily reviewed and pending signature of the Cross-Strait Service Trade Agreement (CSSTA) with China (Figure 1). The occupation of the Legislature would take several weeks and grow into an island-wide movement with strong popular support. In its aftermath it would amend the course of Taiwan politics, as well as its relation to China. It was a major contemporary event in Taiwan, and continues to influence the political landscape and societal reflection in the country. The occupation of the Legislature was streamed live, and when people vacated the chamber they left behind a massive amount of supporting artifacts and documentary materials. What would one do with these artifacts and materials, presumably soon to be abandoned, vanished and forgot? What could an archivist — or anyone who ever was involved in the movement — do in preparing for the future generations to remember the present events?

A few historians in Academia Sinica, Taiwan, seized this opportunity and reached a general agreement with the occupants to systematically collect what was in the Legislature main chamber before they would prepare to end the protest. Afterward Academia Sinica suddenly got hold of a large collection of artifacts created by thousands of participants in a contemporary event. In this paper we discuss a few issues involved in digitizing and archiving artifacts from contemporary events of this nature. We outline our approach to addressing them, and we present the current status of this archive. The archive is called the 318 Civil Movement Archive at Academia Sinica. More background information about it can be found at the archive website [About]. As the movement started on March 18, 2014, has since better known as the Sunflower (Student) Movement [W-en] [W-zh-tw], in this paper the term Sunflower Movement Archive and the term 318 Civil Movement Archive will be used interchangeably to refer to the archive initiated at Academia Sinica. Often we simply call it the Archive.

Principle

To strive for general access to the Sunflower Movement Archive probably was our topmost principle when we were starting to digitize the artifacts. This principle, however, shall be applied to a conflicting context of requirements and constraints. On the one hand, making the archive publicly accessible — on the Web of course — keeps Academia Sinica accountable to the activists (and to the public as well) about what it is doing. Academia Sinica will keep its promise in preserving all the artifacts it has acquired, and the proof is in the form of a Web catalog of all the digitized artifacts. On the other hand, as the artifacts are made by individuals, and some are of a personal nature (encouraging notes to the activists, for example), the individuals' personal privacy, publicity rights, as well as copyrights can be vulnerable when digital copies of the artifacts are made available for all to exam and use.

Because of these considerations, only thumbnail images of the artifacts are made available on the Web for the Archive. The thumbnails are still useful for artifact identification (more about this later), but they are of no plausible other values. In addition, sensitive information inscribed in the artifacts, such as recognizable signatures and phone numbers, has to be pixelated to prevent misuse. No doubt there are boundary cases challenging our judgments. Often we will rather be safe than be sorry, hence will not release even thumbnail images at all for some artifacts. Still, how shall we deal with a banner with hundreds of signatures, sent in by overseas students to support the occupants? Scrubbing out all the signatures from the digital image of the banner will surly defeat the purpose of such an expression of solidarity. We make it a general rule that if it is a form of public communication, it shall be made public, even if there are personal information (names, signatures, affiliations, etc.) on the artifact.

If what are made available are just thumbnail images and artifact metadata, a Web archive will not be too interesting. As participation to the movement is both personal and collective, we hope people will use the online archive to identify artifacts of their own (creation), and to make available high-resolution images of their artifacts to the public for general reuse. That is, we want the Web archive to be a conduit to help transit a collection of orphaned works into a domain of collective remembrance. A feature is built into the online catalog to allow registered users to identify artifacts of their own. Once identified, the user can choose to release the high-resolution image of the artifact to the public under one of the six Creative Commons Licenses, or more openly to elevate it to the public domain by using the CC0 Public Domain Dedication. Of course the claimant can choose to declare to reserve his/her copyright to the work. In this case, the high-resolution image will not be made public. To facilitate better search into the Archive, each item in the collection is annotated with rich metadata, including a transcription of the text appearing on the artifact (the words in a note, for example). People have used this feature to find and release artifacts of their making in spite of (or because of) the artifacts have been archived (and put online) for this historical event.

At the same time when the physical artifacts were being digitized, we also began to collect "born digital" documentary media such as photo images and audiovisual recordings. At the time of the Sunflower Movement, these media were widely dispersed on media sites (e.g. YouTube), social networks (e.g. Facebook), or Web storage services (e.g. Dropbox). After the event, these media may be removed for various reasons, buried in new materials, or hard to find. Many service providers where these media are hosted often scale down the uploaded originals into low-resolution media, transform them into less desirable formats, and/or strip out all the metadata embedded in the original media (e.g. EXIF data in photos). These tainted media are not for archival purposes. We chased down some of the most well-known citizen media activists who were broadcasting and reporting the events. We acquired batches of original files from them. By going after the original producers, we also get to keep better records of the provenance of the digital media in the collection. Many providers chose to donate the entire collections on their hard drives to the Archive, by using the CC0 Public Domain Dedication.

Use and Identification

We feel it necessary to have a Terms of Use (ToU) for the publicly accessible catalog to the Archive [ToU]. By this, we will be able to communicate clearly to the public the purposes of the Archive, as well as various conditions and considerations in using the catalog. As the catalog is free and open to all, even without registration, to search and browse the Archive, we do not want the ToU to sound discouraging. Still, the catalog is the outcome of a provisional project at a research institution. We cannot really warrant the continuity and accuracy of the catalog and its associated services (in particular when funding was very uncertain in the beginning). Nor should we be held liable for people's use of these services. The ToU keeps users aware, and requests their understanding, of the right of publicity, the right to privacy, and other rights of the individuals whose artifacts are collected — or whose appearances are recorded — in the Archive. We also worry about the Archive being used by the authority as a source of evidences to pursue legal cases. Therefore, specifically in the ToU, we ask all users "not to cause civil or criminal disputes," and "not to commit harassment, threat or other misconducts on any individual."

We now demonstrate how anyone can use the Archive. Let us use as an example the hand-written note shown in Figure 2. This item is a small post-it note written by a student from the Chinese University of Hong Kong [10531]. It is part of a large panel sent in by the students from Hong Kong to support the students occupying the Legislature [12958]. The panel is shown in Figure 3, with all the notes attached to it, as it is in the catalog. Each note attached to the panel has been individually digitized and cataloged; the note in Figure 2 is but one of them on the panel. Figure 4 is a photo of the panel hanging on the wall of the main chamber of Legislature during the occupation. As the hand writing in the note has been transcribed into text and becomes part of the item's metadata, one can search for it in the catalog using a few key phrases. In the note, the student says s/he is from the Department of Social Work (社工系). Using this three-character Chinese phrase, we search and indeed find this item in the catalog, as shown in Figure 5.

Would anyone actually use the catalog to search and identify his or her own artifacts? We asked this question ourselves when deciding to add functionality to the catalog to allow registered users to identify, online, artifacts of their own. We were not sure. But once the functionality is there, and after some publicity about our work on the Archive, some people do start to identify their works and mail us their Copyright Declaration and Release forms [Id]. Figure 6 shows a work of art [18247]. It was identified by its creator using the catalog to the Archive. After the identification, he also releases the art work under a CC BY-NC-SA 3.0 TW license. By identifying it, the work can now be attributed to his name (佐瑪, Zuoma). By releasing his works under a Creative Commons license, he allows us to make available high-resolution images of his works for people to download. Figure 7A shows the template of the Copyright Declaration and Release form. A customized form will be generated automatically once an artifact has been identified by its maker. The PDF form will have included all the necessary information (about both the identifier and the identified item). It need only to be printed out, signed, and mailed in — no stamp required (Figure 7B) — by the identifier to Academia Sinica.

Who is Zuoma, the maker of the art work [18247], and what does he look like? People may ask. We shall know as we happen to meet him in person! In Figure 8, he is holding his own work, now a part of the collection of the National Museum of Taiwan History (NMTH). Since November 2016, all physical artifacts in the Archive had been transferred to the museum by a mutual agreement between Academia Sinica and the NMTH. The photo was taken at a press event on 2016-11-14 at the NMTH, where a one-day conference was held on topics of preserving and archiving artifacts from contemporary events. By building information systems encouraging people to reconnect with artifacts that had been forced to be left behind, we aim to help resurrect and disseminate people's stories of the movement. In this particular case, we did get to learn why and how Zuoma made this and other art works in the Sunflower Movement.

Current Status

The catalog of the Archive has been online since March 2015, roughly one year after the events setting off the Sunflower Movement. So far we have not received any complaint about putting the catalog online. For long-term preservation, Academia Sinica has made arrangement with the National Museum of Taiwan History (NMTH) to transfer the Archive to the Museum. The information systems managing the entire collection of digital media, including high-resolution images of all the artifacts, are developed and released as open source software packages. As such, Academia Sinica and the NMTH can both host the digital archive on the Web. Currently the digital archive is still hosted at Academia Sinica even though all the artifacts had been transferred. The museum by itself has been collecting artifacts from various contemporary events for many years, including those from the Sunflower Movement. What Academia Sinica had collected were from the main chamber of the Legislature. The NMTH collects many more from other sources. There is a tentative plan between Academia Sinica and the NMTH to mutually enrich their digital collections on the Sunflower Movement.

An online recollection of the 318 Civil Movement, drawing from a group of individuals loosely connected to the people working on the Archive, was announced and made public on March 18, 2015, the first anniversary of the events. The recollection is a website expressed as a map of Taiwan covered with images and narratives; these are individual stories told with supporting materials drawn from the Archive or from other sources [Expo]. We imagine any person, any group of individuals, can use this catalog to the Sunflower Movement Archive to tell their stories. Each item in the catalog has a permanent link; anyone can use the links to weave stories about the various events in the movement.

Discussion

Remembrance of contemporary events can be both personal and collective. When artifacts are collected from contemporary events, individual and public access considerations constrained what shall and can be done with the artifacts. We hope we have maintained a balance in setting up the Sunflower Movement Archive. We hope our experience can draw some attention to, and incite more discussion about, the issues that are involved in building archives of contemporary events.

We would like to emphasize that the work on the Sunflower Movement Archive is but one among the many in the field of digital archiving and curation. We opt not to give an overview of the best practices, nor cite the many literature, in this brief paper as we fear we cannot do it properly with the current time and space constraints. A balanced and comprehensive survey of the field will be obligatory when an extended version of this paper is to be prepared. Nevertheless we note that our effort is most related to the many existing works in catching the ephemeral but personal (as compared to those in holding on to the permanent and institutional). We look forward to learning from and further extending the practices in post-disaster remembrance and diasporic recollection.

Acknowledgements

The work on the Sunflower Movement Archive has been funded by Academia Sinica and is a collaboration between the Institute of History and Philology and the Institute of Information Science. The author thank all the collaborators, especially Ming-Chong Hwang who initiated the collection of the artifacts and invited the author to join the project to archive them digitally. We thank Chi-Ling Lai, Shih-Ting Tang, and Chia-Hsun Wang for their enduring hard work on this project. Digital Memory Asia did all the digitization works. Word Gleaner Ltd. built all the information systems. Without their help, this project will not finish in time. We are very grateful to Shu-Jiun Chen for her advice on metadata, and to Shun-Ling Chen for her legal advice. We thank the Research Center for Information Technology Innovation, as well as the Office of Vice President Fan-sen Wang, Academia Sinica, for their support. Under a mutual agreement between Academia Sinica and the National Museum of Taiwan History, the entire collection of the Archive had been transferred to the Museum. We thank the Museum for giving the collection a most proper home.

Figures

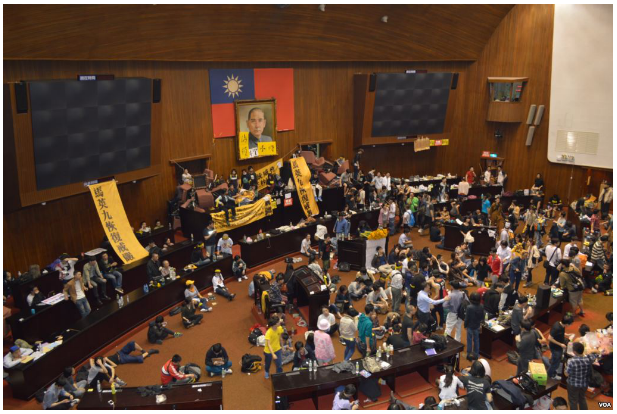
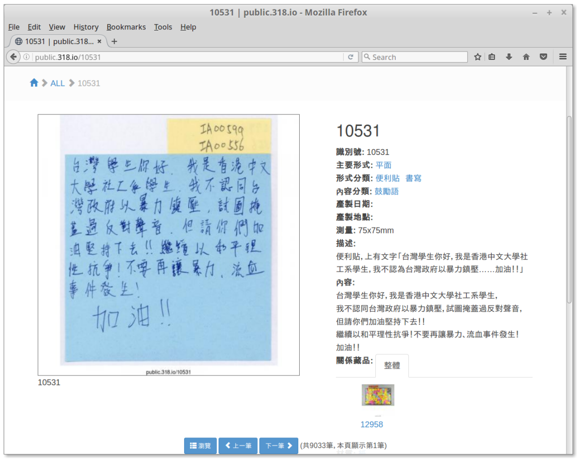


Figure 1. Students occupying the main chamber of the Legislature (photo from the Voice of America [VoA]).

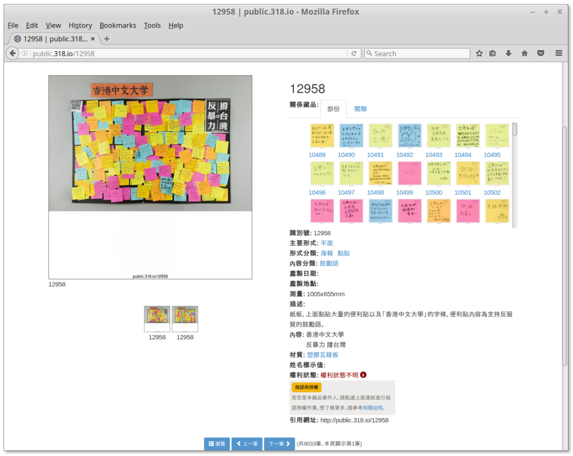


10531

識別號: 10531
主要形式: 平面
形式分類: 便利貼 書寫
內容分類: 鼓勵語
產製日期:
產製地點:
測量: 75x75mm
描述:
便利貼，上有文字「台灣學生你好，我是香港中文大學社工系學生，我不認為台灣政府以暴力鎮壓……加油！！」
內容:
台灣學生你好，我是香港中文大學社工系學生，
我不認同台灣政府以暴力鎮壓，試圖掩蓋過反對聲音，
但請你們加油堅持下去！！
繼續以和平理性抗爭！不要再讓暴力、流血事件發生！
加油！！
關係藏品:

(共9033筆，本頁顯示第1筆)

Figure 2. A supporting note from a student at the Department of Social Work, Chinese University of Hong Kong [10531].



12958

關係藏品:

識別號: 12958
主要形式: 平面
形式分類: 海報 黏貼
內容分類: 鼓勵語
產製日期:
產製地點:
測量: 1005x655mm
描述:
紙板，上面黏貼大量的便利貼以及「香港中文大學」的字條，便利貼內容為支持反服貿的鼓勵語。
內容: 香港中文大學
反暴力 撐台灣
材質: 塑膠瓦稜板
姓名標示值:
權利狀態: 權利狀態不明

若您是本藏品著作人，請點選上面連結進行指認授權作業，想了解更多，請參考相關說明。

引用網址: http://public.318.io/12958

(共9033筆，本頁顯示第1筆)

Figure 3. A panel of supporting notes from the Chinese University of Hong Kong [12958].

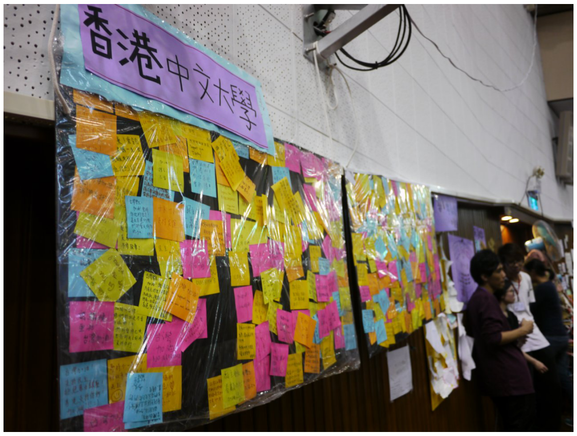


Figure 4. The panel on the wall of the occupied main chamber of the Legislature.

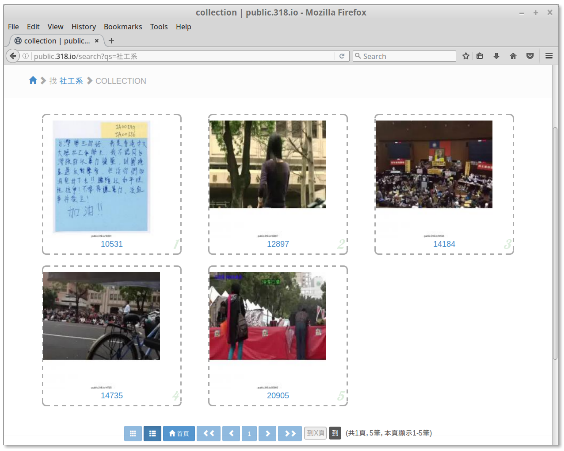
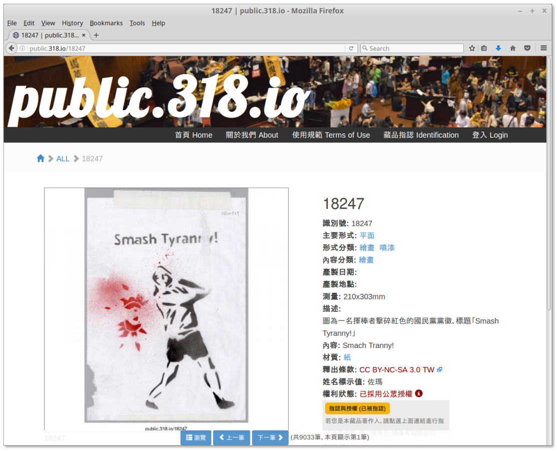

collection | public.318.io - Mozilla Firefox
File Edit View History Bookmarks Tools Help
public.318.io/search?qs=社工系
社工系 > COLLECTION
10531
12897
14184
14735
20905
(共1頁, 5筆, 本頁顯示1-5筆)

Figure 5. The items returned by a search to the Archive with the phrase 社工系 (Department of Social Work).

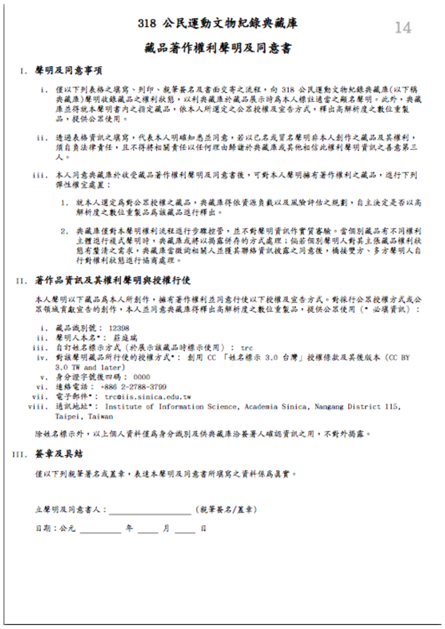


18247

識別號: 18247
主要形式: 平面
形式分類: 繪畫 噴漆
內容分類: 繪畫
產製日期:
產製地點:
測量: 210x303mm
描述:
圖為一名揮棒者擊碎紅色的國民黨黨徽, 標題「Smash Tyranny!」
內容: Smach Tranny!
材質: 紙
釋出條款: CC BY-NC-SA 3.0 TW
姓名標示值: 佐瑪
權利狀態: 已採用公眾授權

若您是本藏品著作人, 請點選上面連結進行指

Figure 6. An identified artifact [18247].



318 公民運動文物紀錄典藏庫

藏品著作權利聲明及同意書

14

I. 聲明及同意事項

i. 僅以下列表格之填寫、列印、親筆簽名及書面交寄之流程，向 318 公民運動文物紀錄典藏庫(以下稱典藏庫)聲明收錄藏品之權利狀態，以利典藏庫於藏品展示時爲本人標註適當之顯名聲明。此外，典藏庫並得就本聲明書內之指定藏品，依本人所選定之公眾授權及宣告方式，釋出高解析度之數位重製品，提供公眾使用。

ii. 透過表格資訊之填寫，代表本人明確知悉並同意，若以己名或冒名聲明非本人創作之藏品及其權利，須自負法律責任，且不得將相關責任以任何理由歸諸於典藏庫或其他相信此權利聲明資訊之善意第三人。

iii. 本人同意典藏庫於收受藏品著作權利聲明及同意書後，可對本人聲明擁有著作權利之藏品，進行下列彈性權宜處置：

1. 就本人選定爲對公眾授權之藏品，典藏庫得依資源負載以及風險評估之規劃，自主決定是否以高解析度之數位重製品爲該藏品進行釋出。

2. 典藏庫僅對本聲明權利流程進行步驟控管，並不對聲明資訊作實質審驗。當個別藏品有不同權利主體進行複式聲明時，典藏庫或將以揭露併存的方式處理；倘若個別聲明人對其主張藏品權利狀態有釐清之需求，典藏庫當徵詢相關人並獲其聯絡資訊披露之同意後，橋接雙方、多方聲明人自行對權利狀態進行協商處理。

II. 著作品資訊及其權利聲明與授權行使

本人聲明以下藏品爲本人所創作，擁有著作權利並同意行使以下授權及宣告方式。對採行公眾授權方式或公眾領域貢獻宣告的創作，本人並同意典藏庫得釋出高解析度之數位重製品，提供公眾使用（* 必填資訊）：

i. 藏品識別號： 12398
ii. 聲明人本名*： 蘇庭瑞
iii. 自訂姓名標示方式（於展示該藏品時標示使用）： trc
iv. 對該聲明藏品所行使的授權方式*： 創用 CC 「姓名標示 3.0 台灣」授權條款及其後版本（CC BY 3.0 TW and later）
v. 身分證字號後四碼： 0000
vi. 連絡電話： +886 2-2788-3799
vii. 電子郵件*： trc@iis.sinica.edu.tw
viii. 通訊地址*： Institute of Information Science, Academia Sinica, Nangang District 115, Taipei, Taiwan

除姓名標示外，以上個人資料僅爲身分識別及供典藏庫洽簽署人確認資訊之用，不對外揭露。

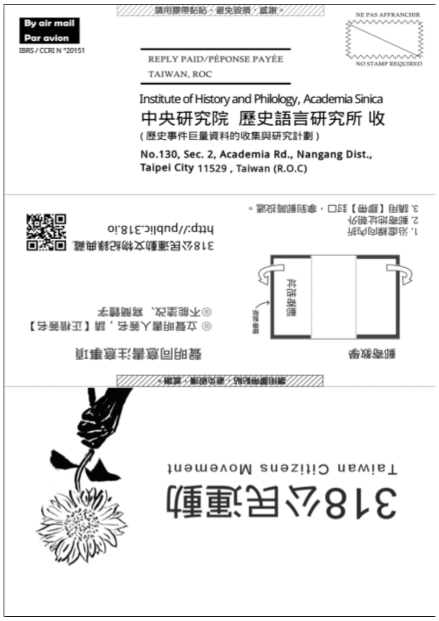
III. 簽章及具結

僅以下列親筆署名或蓋章，表達本聲明及同意書所填寫之資料係爲真實。

立聲明及同意書人：__________________（親筆簽名/蓋章）

日期：公元 _________ 年 ____ 月 ____ 日

Figure 7A. A copyright declaration and release form.



請用膠帶黏貼，避免毀損，感謝。

By air mail
Par avion
IBRS / CCRI N °20151

NE PAS AFFRANCHIR
NO STAMP REQUIRED

REPLY PAID/PÉPONSE PAYÉE
TAIWAN, ROC

Institute of History and Philology, Academia Sinica
中央研究院 歷史語言研究所 收
(歷史事件巨量資料的收集與研究計劃)
No.130, Sec. 2, Academia Rd., Nangang Dist.,
Taipei City 11529 , Taiwan (R.O.C)

318公民運動文物紀錄典藏
http://public.318.io
318公民運動
Taiwan Citizens Movement

Figure 7B. The back cover of the form, to be folded into an envelope — NO STAMP REQUIRED.

為明日收藏：318公民運動與當代典藏
為明日收藏：318公民運動與當代典藏

Figure 8. The entire collection of the Archive was transferred to the National Museum of Taiwan History on 2016-11-14 (as reported by the Liberty Times)

Resources

- [10531] Item no. 10531 in the Archive catalog: <http://public.318.io/10531/>
- [12958] Item no. 12958 in the Archive catalog: <http://public.318.io/12958/>
- [18247] Item no. 18247 in the Archive catalog: <http://public.318.io/18247/>
- [About] Background about the 318 Civil Movement Archive: <http://public.318.io/about/>
- [Expo] Remembrance of the events, as an online group exposition of the Archive catalog: <http://expo.318.io/>
- [Id] Artifacts identified by their makers: http://public.318.io/identified_collections
- [ToU] Terms of Use of the Archive: <http://public.318.io/usage/>.

Bibliography

[LT] **自由時報 (Liberty Times),** 〈318學運文物 台史博接手典藏〉 (Artifacts From The 318 Student Movement Handed Over To The National Museum of Taiwan History), 2016-11-15. <http://news.ltn.com.tw/news/life/paper/1052141>.

[VoA] **美國之音 (Voice of America),** 〈反「兩岸服貿」學生繼續佔據立法院大會堂〉 (Students Against CSSTA Continue To Occupy The Main Chamber Of The Legislature), 2014-03-19. <http://www.voachinese.com/a/taiwan-students-protest-20140319/1874406.html>.

[W-en] **Wikipedia article: Sunflower Student Movement**, accessed 2017-03-31. <https://en.wikipedia.org/wiki/Sunflower_Student_Movement>.

[W-zh-tw] **Wikipedia article: 太陽花學運**, accessed 2017-03-31. <https://zh.wikipedia.org/zh-tw/%E5%A4%AA%E9%99%BD%E8%8A%B1%E5%AD%B8%E9%81%8B>.

Tempi Verbali e Strutture Narrative: l'Analisi Computazionale dei Morfemi Temporali nei Testi Narrativi Italiani tra Realismo e Modernismo

Fabio Ciotti
fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata, Rome

Il problema: la transizione al modernismo nella letteratura Italiana

La transizione dal Realismo al Modernismo, che caratterizzò tutte le grandi letteratura nazionali occidentali nei decenni a cavallo tra l'800 e il 900, è uno dei periodi storico letterari che più hanno attirato l'attenzione della storiografia e della critica letteraria. Si tratta di un fenomeno di portata globale nelle culture occidentali che ha tuttavia avuto specificità stilistiche, scansioni temporali, espressioni di poetica e di produzione testuale diverse in ciascuna tradizione nazionale.

Nella storiografia letteraria italiana la categoria del "Modernismo" ha avuto una fortuna assai scarsa e solo nell'ultimo ventennio essa è entrata nel dibattito teorico, soprattutto per iniziativa di studiosi con prevalente orientamento comparatista (Luperini, 2014; Somigli e Moroni, 2004; Pellini 2004; Guglielmi, 2001). Per lungo tempo infatti si sono preferite nozioni storico letterario come quella di Decadentismo, soprattutto sulla scorta dell'influente lavoro di Salinari (Salinari, 1960) – e in generale di tutta la critica di impianto marxista storicista – secondo cui nel corso dell'ultimo decennio dell'Ottocento comincia a profilarsi, negli ambienti intellettuali italiani, quella "coscienza della crisi", che caratterizzerà la civiltà europea a cavallo tra i due secoli. Una crisi epocale dell'universo ideologico, culturale ed epistemico che aveva dominato nella seconda metà del secolo precedente, e che aveva dato vita alla stagione del Realismo sul piano letterario e del Positivismo su quello filosofico.

Questa periodizzazione e categorizzazione ha ovviamente determinato per lungo tempo notevoli difficoltà nel collocare le opere degli autori principali della letteratura italiana tra i due secoli: *in primis* Pirandello e Svevo, ma anche Tozzi e, in parte almeno, D'Annunzio (Castellana, 2010). Ma l'impianto analitico storicista alla base di tale inquadramento storiografico, concentrato sulla analisi delle poetiche e dei fattori socioculturali e strutturali di contesto della produzione letteraria, ha avuto anche una ulteriore conseguenza: fatta eccezione per alcuni grandi protagonisti, le cui opere sono state analizzate con grande attenzione, di rado si è cercato di verificare quali fossero i tratti testuali della produzione letteraria che permettessero di giustificare le proposte interpretative. La nostra ricerca si è posta dunque come obiettivo la ricerca di quei caratteri testuali intrinseci che fossero in grado di supportare o meno la tradizione interpretativa e storiografica.

Il paradigma metodologico: *distant reading* "critico"

Uno dei temi che hanno ravvivato il dibattito metodologico dell'ultimo decennio in ambito teorico e storico letterario è il paradigma del *distant reading* proposto da Franco Moretti (2013b). La proposta di Moretti, si noti, nella sua formulazione originale non voleva tanto promuovere l'uso di specifici metodi quantitativi computazionali nell'analisi dei testi (ciò che è poi divenuto preponderante nella "vulgata morettiana"), quanto piuttosto richiamare sulla necessità di affrontare una classe di problemi letterari che l'analisi tradizionale non riesce a descrivere correttamente.

Abbiamo più volte espresso le nostre riserve teoriche e metodologiche su alcuni aspetti e applicazioni di questo approccio, riconoscendone tuttavia la validità qualora esso sia applicato su domini e aspetti testuali di adeguato livello descrittivo (per intenderci, fenomeni che comportano l'analisi di corpora testuali vasti per individuare macro-fenomeni sincronici e diacronici), e la sua applicazione sia guidata da un adeguato inquadramento teorico e da specifiche ipotesi interpretative (Ciotti, 2014; 2016).

Crediamo che il problema che ci siamo posti e la sua dimensione cronotopica (la tradizione letteraria Italiana nel periodo che va dalla metà del 1800 al 1920) sia un candidato ottimale per una indagine basata sul *distant reading* e sui connessi metodi computazionali. Nel nostro studio abbiamo circoscritto il dominio dell'analisi alla produzione narrativa (sia nella forma romanzo sia nelle forme brevi) che, per vari rispetti, rappresenta in modo prioritario le problematiche individuate in apertura.

Per quanto riguarda l'inquadramento teorico, siamo convinti che un metodo di analisi computazionale riveste interesse in ambito critico letterario nella misura in cui fornisce dati osservativi che possano essere correlati a termini o nozioni teoriche adottate in una ipotesi interpretativa. Moretti nel suo lavoro più eminentemente metodologico, adotta la nozione di "operazionalizzazione" derivandola dalla epistemologia di P.W. Bridigman (Moretti, 2013a). Non concordiamo con il riduzionismo quantitativo presupposto nella accezione dell'operazionalismo di Moretti. Preferiamo concepire l'analisi computazionale non esclusivamente come un metodo quantitativo/numerico, ma come l'elaborazione di un modello formale funzionalmente isomorfo al dominio, cui si possono applicare processi computazionali. Da punto di vista critico si tratta di una versione computazionale della nozione di interpretazione critico/semiotica proposta da Umberto Eco (1990).

Il metodo: i tempi verbali nel testo e le tesi di Weinrich

Muovendo da queste considerazione metodologiche, il nostro lavoro ha richiesto in prima istanza l'individuazione del quadro teorico di riferimento. Da questo punto di vista ci è sembrato che lo studio del sistema dei tempi verbali secondo le indicazioni a suo tempo fornite da Harald Weinrich (1978) potesse fornire importanti indicazioni per l'analisi narratologica (Cazalé, 1989; Segre, 1985).

Per Weinrich i tempi verbali, nella loro dimensione testuale, non possono essere considerati esclusivamente dei veicoli lessicali atti ad esprimere il "passato", il "presente" e il "futuro" in quanto attributi del tempo reale. A tale visione "referenzialista" lo studioso tedesco oppone una teoria funzionalista dei tempi verbali, i quali, in quanto morfemi ostinati (presenti in gran copia, dunque, in ogni tipo di testo), fanno parte dei segni istruzionali a disposizione dell'emittente per orientare la ricezione del messaggio da lui emesso. I tempi verbali, dunque, appartengono alla classe dei *deittici*, poiché modellizzano la relazione tra il processo di comunicazione e il testo stesso: essi hanno la funzione di istituire e orientare il processo comunicativo, come istruzioni che il lettore deve seguire per recepire la catena sintagmatica correttamente (Weinrich, 1988).

Questo processo di mediazione avviene attraverso tre funzioni fondamentali che caratterizzano il sistema dei tempi verbali. Weinrich le individua in via empirica, basandosi su spogli e analisi di testi narrativi in lingue romanze, in tedesco e in inglese: opposizione tra tempi commentativi e tempi narrativi; divisione dei tempi tra funzione retrospettiva e funzione anticipativa; divisione dei tempi narrativi tra tempi del primo piano e tempi dello sfondo.

La distribuzione paradigmatica e sintagmatica dei tempi verbali nel testo, e le loro reciproche transizioni, dunque, costituiscono la manifestazione sul livello discorsivo del testo del rapporto comunicativo autore-mondo narrativo-lettore, e contribuiscono a manifestare alcuni importanti aspetti strutturali del testo narrativo:

1) rapporto tra autore/narratore e materia della narrazione (eventi e stati narrati);
2) realizzazione discorsiva dei rapporti tra intreccio e fabula;
3) articolazione sintagmatica delle sequenze narrative e descrittive.

Individuare come e in che misura questi aspetti del testo si articolino nel corpus sia a livello diacronico sia a livello sincronico ci permette di indagare il problema della transizione al modernismo nella letteratura italiana partendo da dati testuali, come ci eravamo prefissati.

Naturalmente la fenomenologia di questi elementi della semiotica narrativa non viene esaurita dalla distribuzione di tempi verbali, e a essa contribuiscono in misura notevole gli aspetti semantici della lingua. Si può dire che il sistema temporale dei verbi costituisce un quadro strutturale di fondo che consente al lettore di orientarsi nella ricezione del testo, uno schema rispetto al quale ogni autore costruisce le sue deviazioni idiolettali.

L'analisi

La configurazione quantitativa e sintagmatica dei tempi verbali nei testi è una proprietà testuale lineare che può essere soggetta a scrutinio computazionale con una relativa facilità - sebbene non siano pochi i problemi pratici da affrontare, soprattutto in ragione delle difficoltà che emergono nell'applicazione di sistemi di *part of speech tagging* al rilevamento dei morfemi temporali composti, e in generale a un linguaggio letterario che differisce notevolmente da quello dei *corpora* cui sono abitualmente applicati. Data la natura statistica dell'analisi tuttavia, una dose di errore statistico nel riconoscimento dei morfemi temporali è comunque accettabile.

Il corpus testuale di riferimento (composta da circa 400 testi unici) è stato estratto in gran parte dalla collezione del progetto Bibit. Alcune ulteriori edizioni digitali sono state estratte da *corpora* minori, tra cui le collezioni testuali prodotte dall'autore nel corso della sua pregressa attività didattica e di ricerca.

I risultati ottenuti finora sembrano indicare che esiste una mutazione nella configurazione dei morfemi temporali coincidente con la transizione di fase letteraria. Questa a sua volta è funzionale a una destrutturazione dell'impianto narrativo realista che avviene sia mediante una rimodulazione dei canoni tematici, sia mediante la trasformazione delle strutture formali e compositive dell'intreccio e del rapporto tra voce narrante/punto di vista e situazione narrativa. La transizione di paradigma culturale di fine secolo si riflette dunque all'interno dei sistemi modellizzanti secondari (Lotman, 1990), come quello letterario, nell'abbattimento dei confini tra i generi, nella commistione tra stili e registri espressivi e nella rifunzionalizzazione delle strutture narrative.

Bibliografia

Castellana, R. (2010). Realismo modernista. Un'idea del romanzo italiano (1915-1925). *Italianistica*, 1: 23-4

Cazalé, C. (1989). Tempo, Azione, Identità: costanti narrative nella raccolta Scialle Nero. *Rivista di studi pirandelliani*, 2 (III s.): 81-101

Ciotti, F. (2014). Digital Literary and Cultural Studies: State of the Art and Perspectives. *Between*, 4(8). doi:10.13125/2039-6597/1392. http://dx.doi.org/10.13125/2039-6597/1392.

Ciotti, F. (2016). What's in a Topic Model. I fondamenti del text mining negli studi letterari. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 149-151.

Eco, U. (1990). *I limiti dell'interpretazione*. Milano: Bompiani.

Guglielmi, G. (2001). *L'invenzione della letteratura. Fra modernismo e avanguardia.*
Napoli: Liguori.

Lotman J. (1990). *La struttura del testo poetico.* Milano: Mursia.

Luperini, R. (2014). Modernismo, avanguardie, antimodernismo. *Laletteraturaenoi*, http://www.laletteraturaenoi.it/index.php/interpretazione-e-noi/271-modernismo,-avanguardie,-antimodernismo.html.

Moretti, F. (2013a). Operationalizing: Or, the Function of Measurement in Literary Theory. *New Left Review*, 84: 103-119.

Moretti, F. (2013b). *Distant Reading*. London: Verso.

Pellini, P. (2004). *In una casa di vetro.* Firenze: Le Monnier.

Salinari C. (1960). *Miti e coscienza del decadentismo europeo*, Milano: Feltrinelli.

Segre, C. (1985). *Avviamento all'analisi del testo letterario.* Torino: Einaudi.

Une Preuve de concept pour la sémantisation et la visualisation orientée utilisateur de données archivistiques

Florence Clavaud
florence.clavaud@culture.gouv.fr
Archives Nationales, France

Emmanuel Château-Dutier
emmanuel.chateau.dutier@umontreal.ca
Université de Montréal, Canada

Notre présentation concerne un travail exploratoire aux objectifs qualitatifs, qui ambitionne de fournir des pistes de réflexion pour avancer vers la sémantisation des métadonnées des documents d'archives et la représentation à l'écran de ces données RDF, en vues multidimensionnelles, porteuses de sens et manipulables.

Trois organismes français (les Archives nationales de France, la Bibliothèque nationale de France, le Service interministériel des Archives de France) et un laboratoire de recherche canadien (le Département d'histoire de l'art et des études cinématographiques de l'Université de Montréal) se sont associés en 2015 pour réaliser une preuve de concept visant à démontrer qu'il est possible :

- de représenter en RDF, en veillant à la précision, à l'exactitude et à l'utilisabilité des triplets obtenus, des métadonnées archivistiques produites de différentes manières et selon diverses perspectives (celles d'institutions patrimoniales, celles d'un chercheur) ;
- d'enrichir les triplets obtenus en créant de nouveaux triplets, qu'il s'agisse de procéder à des alignements ou d'établir de nouvelles relations par inférence ;
- de produire une interface de recherche et d'exploration analytique et graphique qui soit dynamique, ergonomique et signifiante, sans sacrifier la granularité informationnelle ni la lisibilité.

Il s'agit donc de réaliser un démonstrateur, sous la forme d'une application web construite en utilisant des composants libres, dont les sources seront placées sous licence libre et déposées dans un entrepôt public.

Cette opération est relativement complexe. Il n'y a pas de réel précédent dans le domaine des archives. De plus, jusqu'à ce jour, il n'existait pas d'ontologie générique du domaine. Enfin, il n'existe pas a priori de librairie ou logiciel satisfaisant directement la totalité des besoins, en particulier pour ce qui concerne la visualisation des jeux de données.

Les bibliothèques, comme la Bibliothèque nationale de France (avec *data.bnf.fr),* et plusieurs musées, se sont lancés dans la conversion en RDF de leurs métadonnées et dans la réalisation d'interfaces de présentation de ces données, dès lors que des modèles conceptuels et ontologies appropriés ont été disponibles pour représenter leurs collections (CIDOC-CRM et FRBR). Par contre, si divers projets spécifiques ont déjà, soit produit des ontologies pour la description des archives et mis en œuvre ces ontologies (par exemple le projet LOCAH, dont les résultats sont consultables via *The Archives Hub Linked Data*), soit cherché à produire des représentations analytiques ou graphiques de ces descriptions, jusqu'ici ces réalisations n'avaient pas en même temps les ambitions de la généricité, de la complétude et d'une grande précision.

Il a fallu attendre fin 2016 pour la publication d'un modèle conceptuel global pour la description des archives (*Records In Contexts-Conceptual Model* ou RiC-CM) ; l'ontologie OWL correspondante, *Records In Contexts-Ontology* ou RiC-O), qui est l'ontologie de référence pour notre projet, sera publiée en 2017.

En outre, si la data visualisation est aujourd'hui devenue, en s'appuyant sur des technologies de plus en plus performantes, un important domaine d'action et d'innovation pour les humanités numériques, il est à notre connaissance encore très rare de trouver articulés la dimension temporelle, très forte en archivistique et en histoire, et une représentation précise, en graphe, d'objets historiques et des différents éléments de leur contexte, le tout étant susceptible de former un réseau très dense. Ainsi, les divers projets réalisés jusqu'ici présentent uniquement de façon relativement simple les agents et les documents (bibliothèque publique de la ville de New York), ou encore uniquement des réseaux d'agents (SNAC), ou encore des agents, leur histoire et leurs relations dans une perspective diachronique, mais pas les ressources documentaires liées (*Kindred Britain*).

Dans le cas de notre projet, nous avons commencé par choisir un périmètre pour les jeux de métadonnées source et préparer ces jeux de métadonnées, chaque partenaire travaillant de son côté, selon sa propre perspective, tout en respectant quelques règles élaborées en commun.

Les jeux de métadonnées sont composés de notices descriptives d'organismes et personnes physiques acteurs dans deux domaines fonctionnels de l'administration française (la gestion des monuments historiques et bâtiments civils des années 1795 à nos jours ; la gestion des bibliothèques publiques et de la lecture du 19e siècle à nos jours) et d'instruments de recherche archivistiques décrivant les ensembles documentaires produits par ces entités. Ils sont conformes aux standards archivistiques actuels que sont les normes ISAD(G) et ISAAR(CPF) et leurs transpositions techniques, la DTD EAD 2002 et le schéma XML/EAC-CPF. Nous avons choisi de procéder nous-mêmes à la sémantisation de ces jeux de métadonnées. Un vocabulaire des fonctions des entités encodé en SKOS/RDF, le seul à avoir été produit collectivement, s'ajoute à ce corpus.

Un cahier des charges a été élaboré, une consultation lancée et une société de services a été choisie fin octobre 2016 pour concevoir et réaliser le démonstrateur. Elle réalisera le travail selon une approche agile, en interaction forte avec l'équipe projet, qu'il s'agisse d'analyser l'ontologie et les triplets RDF fournis en entrée, de bâtir des scénarios de recherche, d'alignement, d'enrichissement ou de visualisation, ou de tester les versions successives du logiciel.

Comme le projet sera achevé fin octobre 2017, une version quasiment définitive du prototype devrait être disponible en août 2017. Un bilan critique, méthodologique et prospectif sera ensuite réalisé et publié par les entités impliquées. En avant-première, après une présentation rapide des objectifs, des enjeux et de l'historique du projet, nous nous attacherons à évoquer deux de ses aspects principaux :

- la mise en œuvre, pour la représentation en RDF des métadonnées archivistiques retenues, de l'ontologie RiC-O : présentation de l'ontologie et de ses principes de conception, discussion des choix d'adaptation qui ont été faits pour les besoins du projet, présentation des résultats obtenus, évaluation de ces résultats en ce qui concerne la granularité d'expression et les possibilités de raisonnement induites ;
- la conception de l'interface de recherche et de visualisation : présentation de la méthode suivie, des choix faits et des résultats obtenus, de leur intérêt et de leurs limites pour l'utilisateur final, que celui-ci soit un chercheur averti ou un amateur moins connaisseur des concepts archivistiques.

Bibliographie

Bertin, J. (1977). *La graphique et le traitement graphique de l'information.* Paris : Flammarion.

Bibliothèque nationale de France (2011-). *Data.bnf.fr.* http://data.bnf.fr

Drucker, J. (2014*). Graphesis: visual forms of knowledge production.* MetaLABprojects. ISBN 9780674724938

Fekete, J-D. (2010). "Visualiser l'information pour la comprendre vite et bien". *Dans L'Usager numérique*. Séminaire INRIA, 27 septembre-1er octobre 2010. ADBS, pp. 161-194.

Few, S. (2013). " Data Visualization for Human Perception ". Dans Soegaard, M. et Friis Dam, R. (éds.), *The Encyclopedia of Human-Computer Interaction,* 2e édition. En ligne : https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception

Friendly, M. (2009). *Milestones in the history of thematic cartography, statistical graphics, and data visualization.*Version PDF en ligne :

http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf

Gibbs, F. W. (2016). “ New Forms of History: Critiquing Data and Its Representations ”. *The American Historian,* 7.En ligne : http://tah.oah.org/february-2016/new-forms-of-history-critiquing-data-and-its-representations/

Grandjean, M. (2014). “ La connaissance est un réseau ”. *Les Cahiers du Numérique,* 10/3 : 37–54. DOI :10.3166/LCN.10.3.37-54.

Grandjean, M. (2015). “ Introduction à la visualisation de données, l’analyse de réseau en histoire ”. *Geschichte und Informatik,* 18-19 : 109-128. Version PDF en ligne : http://www.martingrandjean.ch/wp-content/uploads/2015/09/Grandjean2015.pdf

Januals, B. et Minel, J-L. (2016). “ La construction d’un espace patrimonial partagé dans le Web de données ouvert ”. *Communication*, 34/1. En ligne : http://communication.revues.org/6650. DOI : 10.4000/communication.6650.

Jenkins, N., Meeks, E. et Murray, S. (2013). *Kindred Britain.* Stanford University Libraries. http://kindred.stanford.edu/

Manovich, L. (2010). What is Visualization? En ligne : http://manovich.net/content/04-projects/064-what-is-visualization/61_article_2010.pdf

Natale, E., Sibille, Ch., Chachereau, N., Kammerer, P., Hiestand, M. et al (2015). “ La visualisation des données en histoire / Visualisierung von Daten in der Geschichtswissenschaft ”. *Geschichte und Informatik*, 18/19.

Krempel, L. (2011). “ Network visualization ”. Dans Scott, John G. et Carrington, P. J. (éds.), *The SAGE handbook of social network analysis*. Pp. 558-577.

International Council on Archives. Experts Group on Archival Description (2016). *Records In Contexts: a conceptual model for archival description*. Consultation Draft v0.1. En ligne : http://www.ica.org/fr/publication-de-records-contexts-par-legad

International Council on Museums (ICOM). International Committee for Documentation (CIDOC) (2011). CIDOC Conceptual Reference Model. Version 5.0.4. En ligne : http://www.cidoc-crm.org/get-last-official-release

International Federation of Libraries Associations and Institutions (2008). *Functional Requirements for Bibliographic Records* (FRBR). En ligne : http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf (en) ; http://www.bnf.fr/documents/frbr_rapport_final.pdf (fr).

International Federation of Libraries Associations and Institutions (2016). *FRBR-Library Reference Model (FRBR-LRM).* Draft for world-wide review. En ligne : http://www.ifla.org/files/assets/cataloguing/frbr-lrm/frbr-lrm_20160225.pdf

International Working Group on FRBR and CIDOC CRM Harmonisation (2008). *FRBR object-oriented definition and mapping to FRBRer.* Version 0.9 draft. En ligne http://archive.ifla.org/VII/s13/wgfrbr/FRBRoo_V9.1_PR.pdf .

The New York Public Library. (n. d.) Archives and manuscripts: beta tools & experiments. http://archives.nypl.org/tools

University of Liverpool (2010-2013). *The Archives* hub Linked Data. http://data.archiveshub.ac.uk/

University of Virginia Institute for Advanced Technology in the Humanities. (n.d.) *Social Networks and Archival Contexts* (SNAC) prototype. http://socialarchive.iath.virginia.edu/home_prototype.html

Yau, N. (2013). *Data Points: Visualization That Means Something.* John Wiley & Sons. ISBN 978-1-118-46219-5

Quill: Reconstructing the Secretary's Desk for the Records of the 1787 Convention

Nicholas Cole
nicholas.cole@history.ox.ac.uk
Univeristy of Oxford, United Kingdom

Alfie Abdul-Rahman
alfie.abdulrahman@oerc.ox.ac.uk
Univeristy of Oxford, United Kingdom

Grace Mallon
grace.mallon@univ.ox.ac.uk
Univeristy of Oxford, United Kingdom

Kate Howarth
kate.howarth@pmb.ox.ac.uk
Univeristy of Oxford, United Kingdom

Introduction

The standard process for negotiating legal and quasi-legal texts over more than two hundred years has been a parliamentary one that (with variations) is still recognizable as the one described in Thomas Jefferson’s *Manual of Parliamentary Practice* (1801) (U.S. Government Publishing Office, 2016; May, 1844). Proposals are examined by a series of committees, amendments being proposed and voted on throughout this process. Since the late eighteenth-century, many of these negotiations of historical note have left records that record the proposals made and the outcome of decisions taken. Such records are difficult to read—especially when they concern any protracted or complicated process of negotiation, since it rapidly becomes impossible for a reader to keep track of the state of the documents under discussion. Fully comprehending the records if not read in chronological sequence is impossible.

Building on code written for collaborative document editing, we have built a sophisticated, web-accessible platform for the study of negotiated texts. We kept the underlying data-model as simple and generic as it could be while modelling the various procedures suggested by a range of Parliamentary Procedure handbooks. We considered the needs of several distinct classes of users — those doing the work of data-capture, those reviewing that work, those wishing to comment on the detail of the text, producing secondary materials for a variety of audiences, and those wishing to navigate through the material for a variety of purposes.

Initial application

One such process of negotiation was that which created the United States Constitution. The Records of the 1787 Convention, despite being imperfect and not (initially) intended for public release, in fact enable a detailed reconstruction of the work of the Constitutional Convention. These records have been available in various printed forms since 1819 when the official *Journal* was first printed; these printed records have been digitized as both images and transcribed text (Lilian Goldman Law Library, 2008; Silverbrook and Johnson, 2007; National Archives Catalog, 2016; Library of Congress, 2016). While indexing and searching increase the utility of both paper and digitized versions (by allowing readers to discover when particular topics were debated), neither format allows the reader to understand the full context of a particular debate. This is of no small importance, because opinions of participants in the negotiations about particular matters shifted as surrounding questions were answered one way or another.

Other digital projects

Whereas the *Comparative Constitutions Project* has pioneered the comparison and display of finished constitutional texts (Elkins and Ginsburg, 2005), and some other web projects have attempted to make the text of the United States constitution easier to navigate (Surden, 2015), our project focuses instead on the process of negotiation. Other projects have attempted to overcome the limitations of the Convention's records by giving users narrative outlines of the key events in a way that can guide their reading (Lloyd, 2016; Linder, 2016; EDSITEment, 2016). More generally, websites tracking the progress of Parliamentary debates have focused on **milestone** moments in the history of texts, rather than letting users track the detail of a document's evolution (Parliament, 2016; Tauberer, 2004).

Challenges

The records relating to formal negotiations are typically a set of minutes that record of proposals and the votes taken upon them. The principal aim of those recording the minutes is to facilitate the record-keeping process necessary during the work of committees, not to provide later readers with an easy way to reconstruct any particular moment. Each formal proposal to amend a document has at least two contexts that are relevant to readers — what does the document look like when the amendment is proposed? what does the document look like when the amendment is approved or rejected? Due to the nature of committee work, these contexts may differ significantly. Making sense of these records, therefore, poses a significant memory-burden on readers. Detailed and specific discussion of issues presented by these records is hampered by the need for authors to provide their own reconstruction of elements of this process, presented in a narrative form that is necessarily partial and can itself become difficult for readers to follow.

This problem might have been partially addressed using creative re-purposing of either version-control systems designed for computer-science applications (such as the tools *rcs* (GNU, 2013), *git* (Software Freedom Conservancy, 2016), or *mercurial* (Mercurial, 2016)) or else the creation of a layered XHTML document (The University of Virginia Press, 2009). However, we rejected these solutions as being either incapable of fully capturing the nature of the source material or else as likely to result in a fragile platform that would have been too much tied to the specific documents and unsuitable for more general applications. Since future work will compare different negotiations, a more generic platform that could work with a variety of sources with minimal new coding was required.

Our solution

Quill is a newly developed platform for the study and presentation of formal negotiations. It was developed initially with a view to presenting the records of the 1787 Federal Convention that created the Constitution of the United States, but was designed to be a generic platform applicable to a wide range of materials, including the creation of constitutions, treaties, and legislation. The model captures formal negotiations — that is those where there is a procedure of considering and deciding upon discrete suggestions for the wording of documents, and where minutes capturing these deliberations have been taken.

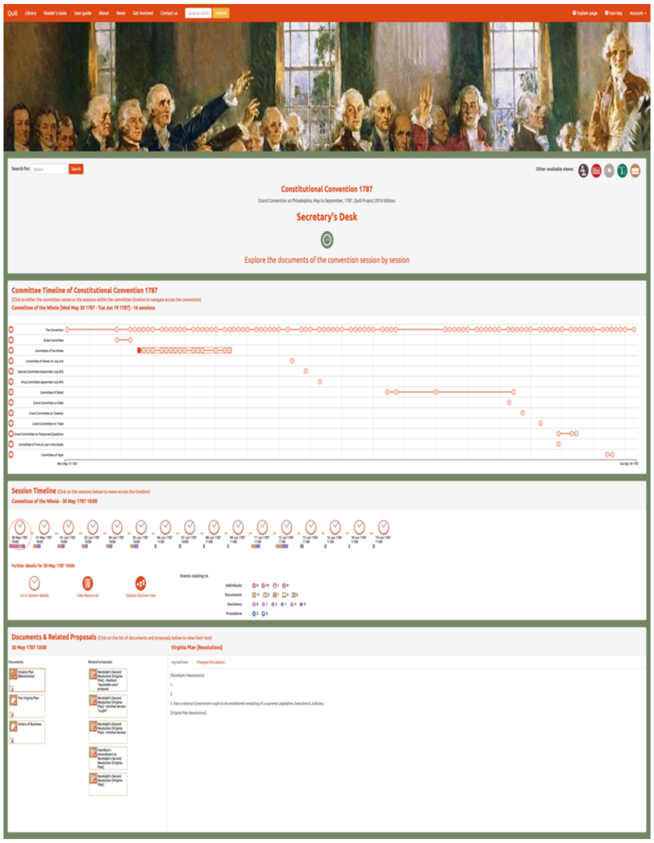
An innovation was to present not merely the reconstruction itself but to integrate a publishing platform that would allow authors to present their own commentary on the material in a way that would allow analysis to be presented alongside specific events within the timeline.

Links to relevant material held on other websites (for example, images of the original manuscripts) are similarly presented to users where relevant. In this way, the website integrates with existing materials, enhancing their value as well as its own and avoiding duplication of effort. We encourage such co-operation with other projects through machine-readable interfaces, a flexible permissions system and a system of **resource collections** that allow third-parties to manage links to their own assets and control how they appear within our platform.

For the 2016 release we relied on the 1911 edition of the Convention Records published by Farrand, even though we know these to be imperfect transcriptions of the surviving manuscripts. This choice allowed us to focus on the development of the software platform. The Quill Platform is capable of presenting different versions of the same event, and future work using the original manuscripts will refine our presentation of the records.

Supporting information–seeking and exploration

Negotiations of this type are extremely complex and assisting users to access the information they require is a challenge. In our public interfaces, we have guided users to access material in several ways. To acclimatize users to the idea of navigating the history of an evolving text, we present a *Secretary's Desk* view, which allows users to navigate the state of documents as they existed at the end of each committee session (see Figure 1).



Constitutional Convention 1787

Grand Convention at Philadelphia, May to September, 1787, Quill Project 2016 Edition.

Secretary's Desk

Explore the documents of the convention session by session

Committee Timeline of Constitutional Convention 1787

(Click on either the committee names or the sessions within the committee timeline to navigate across the convention)

Committee of the Whole [Wed May 30 1787 - Tue Jun 19 1787] - 16 sessions

Session Timeline (Click on the sessions below to move across the timeline)

Committee of the Whole - 30 May 1787 10:00

Further details for 30 May 1787 10:00:

Documents & Related Proposals

Virginia Plan [Resolutions]

Figure 1: The Secretary's Desk for the Committee of the Whole on the 30 May 1787

This view hides much of the complexity of the negotiations, but allows new users to quickly grasp the concept of our reconstruction. We also present visualizations that allow users to explore the structure of negotiations through a variety of tree-diagrams (Herman, *et al.* 2000) (see Figure 2) and sunburst-style (Stasko, *et al.* 2000) visualizations.

The role of individuals and specific delegations, as well as voting patterns, are presented in separate visualizations. All of these views guide users who need more detail down to views that present the work of individual committee sessions moment by moment. Users looking for information on a specific topic are guided towards a search tool. In addition, users can also navigate the platform through a variety of resource and commentary collections, making it possible to provide users with a more guided experience.

Conclusions and future work

The process of negotiating the constitution was complicated (we have modelled close to 4,000 discrete events), and our presentation transcends the possibilities of narrative accounts while making access to and intelligibility of the extant sources much quicker and of greater utility for a broad range of users. The model system itself is content-agnostic and could be used to model a wide range of similar processes. Future work will continue to enhance the user experience both through refinement of the visualizations and user interface, and through the creation of guided views in to the material in collaboration with others, as well as expanding the range of material.

Wider public engagement and education is a key aim of this project. We are collaborating with non-profit organizations in the United States to develop guided views suitable for classroom use and integrated in to existing curriculum materials.

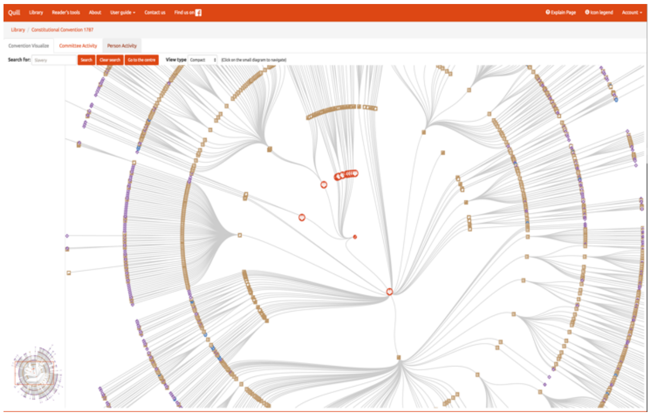


Figure 2: An Activity View showing the work of the whole 1787 Convention showing the hierarchical relationship between events, documents, and committees

Bibliography

EDSITEment, (2016). The Constitutional Convention of 1787. Available at https://edsitement.neh.gov/curriculum-unit/constitutional-convention-1787. Date last accessed: 1 Nov 2016.

Elkins, Z. and Ginsburg, T. (2005). Comparative Constitutions Project. Available at http://comparativeconstitutionsproject.org/. Date last accessed: 1 Nov 2016.

GNU, (2013). GNU RCS. Available at http://www.gnu.org/software/rcs/rcs.html. Date last accessed: 1 Nov 2016.

Herman, I., Melançon, G., and Marshall, M. S. (2000). "Graph Visualization and Navigation in Information Visualization: A Survey". *IEEE Transactions on Visualization and Computer Graphics*, 6(1), pp. 24-43.

Library of Congress, (2016). Part of: James Madison Papers, 1723-1859. Available at https://www.loc.gov/search/?fa=partof:james+madison+papers,+1723-1859:++subseries+5e,+madison%27s+original+notes+on+debates+in+the+federal+constitutional+convention,+1787. Date last accessed: 1 Nov 2016.

Lilian Goldman Law Library, (2008). Notes on the Debates in the Federal Convention. Available at http://avalon.law.yale.edu/subject_menus/debcont.asp. Date last accessed: 1 Nov 2016.

Linder, D. (2016). The Constitutional Convention 1787. Available at law2.umkc.edu/faculty/projects/Ftrials/conlaw/convention1787.html. Date last accessed: 1 Nov 2016.

Lloyd, G. (2016). The Constitutional Convention. Available at teachingamericanhistory.org/convention/themes/. Date last accessed: 1 Nov 2016.

May, T. E. (1844). "A Treatise upon the Law, Privileges, Proceedings and Usage of Parliament". Charles Knight & Co., London, United Kingdom.

Mercurial, (2016). Mercurial. Available at: https://www.mercurial-scm.org. Date last accessed: 1 Nov 2016.

National Archives Catalog, (2016). Journal of the Federal Convention. Available at https://catalog.archives.gov/id/7347105. Date last accessed: 1 Nov 2016.

Parliament, (2016). Bills before Parliament 2016-17. Available at http://services.parliament.uk/bills/. Date last accessed: 1 Nov 2016.

Silverbrook, J. and Johnson, M, (2007). ConSource. Available at http://www.consource.org/. Date last accessed: 1 Nov 2016.

Software Freedom Conservancy, (2016). Git. Available at https://git-scm.com. Date last accessed: 1 Nov 2016.

Stasko, J., Catrambone, R., Guzdial, M., and McDonald, K. (2000). "An Evaluation of Space-filling Information Visualizations for Depicting Hierarchical Structures". *International Journal of Human-Computer Studies*, 53(5), pp. 663-694.

Surden, H. (2015). Constitution Explorer (Beta). Available at http://www.harrysurden.com/projects/visual/US-Code_D3/constitution/US_Constitution_Tree.html. Date last accessed: 1 Nov 2016.

Tauberer, J. (2004). Bills and Resolutions. Available at https://www.govtrack.us/congress/bills/. Date last accessed: 1 Nov 2016.

The University of Virginia Press, (2009). Herman Melville's *Typee*: A Fluid-Text Edition. Available at http://rotunda.upress.virginia.edu/melville/default.xqy. Date last accessed: 1 Nov 2016.

U.S. Government Publishing Office, (2016). House Rules and Manual. Available at https://www.gpo.gov/fdsys/browse/collection.action?collectionCode=HMAN. Date last accessed: 1 Nov 2016.

Getting at Metaphor

Katharine Coles
katharine.coles@utah.edu
University of Utah, United States of America

This paper will discuss how our newly prototyped POEMAGE visualization tool (see McCurdy et al, 2015), created to identify and visualize complex sonic relationships within individual poems, has provided poetry scholars with new ways to identify and conceptualize metaphor, which has previously been considered computationally intractable because of its semantic and syntactic complexities. It focuses not on the tool's technical details but on the ongoing re-theorization of poetry it has engendered,

Close readers are trained to connect every element in a poem to every other in an ambiguous, shifting complex of meaning, which the reader, bringing her own complexities to the process, activates. This poetic dynamic makes computational analysis and visualization of any aspect of poetry a challenge. The goal of our research team has been to take a single poetic element—sound—and treat it computationally and visually at a level of complexity that will make POEMAGE useful to poets and scholars performing sophisticated close readings of poetry, even as it makes poetry more accessible to students and casual readers. Though sound interacts with the other features operating within a poem, unlike most other features it can be looked at in its own terms and is subject to computer analysis through quantification.

As we began, poet Julie Gonnering Lein and I sought to preserve poetry's qualitative, aesthetic experience; computer scientists Miriah Meyer and Nina McCurdy sought to address open questions in their field. Both goals required moving beyond what the machine could already do. Off-the-shelf software can see exact rhyme quickly, as can a good reader—who will swiftly move on to look for sonic relationships that don't replicate themselves but enact disruptive changes that are hard to identify computationally. To capture the progression of sonic clusters as they repeat in different and evolving combinations not only within but across syllables presents a computational problem that required our technical team to develop RhymeDesign, which allows users to query a broad range of sonic patterns within a poem and to design custom templates to query patterns we haven't imagined. Built on top of RhymeDesign, the POEMAGE interface visualizes and allows users to explore interactions of the queried patterns.

In performing this work, we have looked for (and not yet found) computational breakthroughs that might bring metaphor within reach, a process that has required us to consider closely how metaphor works. The difficulty with metaphor inheres even in simple instances. Getting the machine to understand why "Hope is a bird" (or, more problematically, "'Hope' is the thing with feathers") is a metaphor but "Juliet is a Capulet" and "Karen is a Carpenter" may be either similar or different statements of fact is not straightforward.

Poets as different as Dickinson and Donne play complex metaphors out across entire poems in elaborate and shifting figural structures. To develop a tool that can reliably identify metaphoric relationships as POEMAGE identifies sonic relationships—in real time across the entire poetic field—would require the solution of multiple open problems in computer science.

However, recent readings of poems by Dickinson and others, undertaken using POEMAGE, suggest that it is possible to use the tool to access some metaphors not directly but indirectly, leveraging the fact that both rhyme and metaphor operate by substituting one word for another that is different-but-similar, and that inevitably sites of sonic difference-in-similarity point to semantic difference-in-similarity as well. In close reading, we have noted that places the machine marks as being sonically "interesting" are also sites of metaphorical action, and that this action often inheres in, rather than simply existing alongside, the sonic relationships being indicated by the tool. This inherence can emerge through various kinds of sonic relationships, including but not limited to homonyms like "knot" and "naught," which POEMAGE shows in Bradstreet's "Prologue," and eye-rhymes like "blood" and "mood," which it picks up across Pelizzon's "Blood Memory," about menstruation. In presenting these words as related, even conjoined, the machine opens a space for us to tease out figural connections between a loop in a rope and nothingness, or menstrual blood and emotional pain.

A more complex example of metaphor developing through sound occurs in Dickinson's #313:

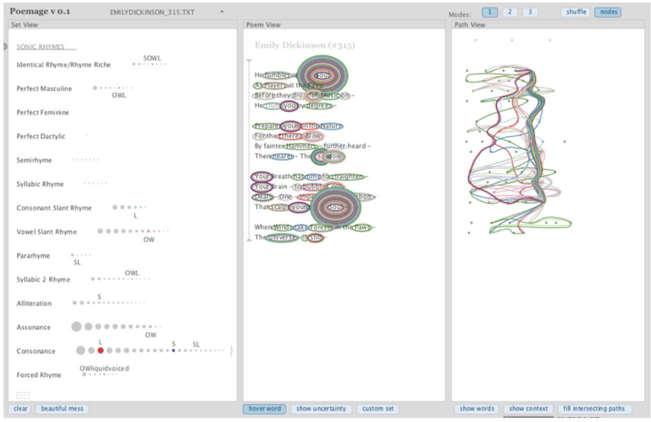
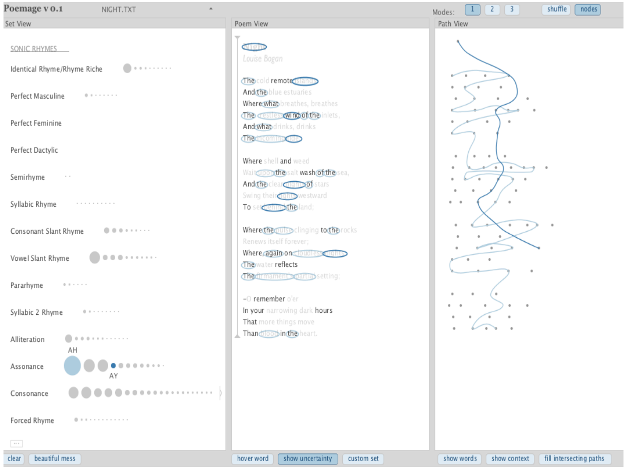


Figure 1: The user selects from the sonic rhymes available in the left-hand panel. The middle panel shows by color which words are implicated in a particular rhyme and which share phenomes. The right-hand panel shows how various rhymes flow through the poem.

The visualization, which shows words that connect with "soul" through specific sonic relationships, makes clear at a glance that "soul" is sonically implicated with virtually every other word in the poem, though of course it is more strongly associated with some words than others. Its identification with "you" and "your" is notable, but I am more attentive to its strong links to "so" (a perfect rhyme and identical except for one dropped phoneme), "slow" (which deploys exactly the same three phonemes, with the second two reversed), and "still," which begins and ends with the same phoneme. Here, then, the "soul," identified with "you," is also sonically and so (because words mean) semantically identified with, in this order, intensity, slowness, and stillness. Beyond this, the tool invites an unlikely leap: the connection of "soul" with the poem's last word, "paws"—a big enough stretch that I am not sure I would make it without the tool's suggestion. This sonnet, which ends with what should be a rhymed couplet, teases by failing to do so. However, the tool's connection of "paws" with "slow" and, through "slow," to "soul" and finally "still," may suggest to the attentive reader an absent but implied homonym, "pause." Here, through an indirection of sound, the poem creates a semantic "rhyme" between "paws/pause" and "still."

Another case arises from our group's interest in uncertainty analysis, rooted in our understanding of ambiguity as a fundamental function of language, which led us to include in POEMAGE a "show uncertainty" button. Even beyond vagaries of accent, numerous words—*tear* comes to mind immediately—can be pronounced in more than one way, and mean differently depending on pronunciation. Every such word requires the machine to make a "guess"—a statistical prediction—about how the word is pronounced within the poem. The "show uncertainty" button allows the user to see words with alternate pronunciations as well as how the machine has chosen to pronounce them.

In "Night" by Louise Bogan, the tool mishears [short i] "wind" as [long i] "wind" (note that it rhymes with "tide":



Night
Louise Bogan

The cold remote islands
And the blue estuaries
Where what breathes, breathes
The restless wind of the inlets,
And what drinks, drinks
The incoming tide;

Where shell and weed
Wait upon the salt wash of the sea,
And the clear nights of stars
Swing their lights westward
To set behind the land;

Where the pulse clinging to the rocks
Renews itself forever;
Where, again on cloudless nights,
The water reflects
The firmament's partial setting;

–O remember o'er
In your narrowing dark hours
That more things move
Than blood in the heart.

Figure 2

Given its instructions, and the overwhelming probability that any one-syllable word comprising a beginning consonant and "ind" will have a long and not a short vowel, the computer had little choice but to make the "judgment" it did, even though the proficient *human* native speaker would not make this delightful mistake. In this "mishearing" of the poem, an inlet winding becomes "restless," and in this restlessness the central metaphor of the poem takes shape. Through such reading, we can see that tools designed to help forestall error in science may reveal disturbances that add to the richness of a poem by opening interpretive space and not only the possibility for but the actual presence of metaphor.

Of course, "getting at metaphor" in this oblique way, even if we give it a separate button in the tool, is not the same as creating a tool that will algorithmically identify and visualize metaphor through its syntactic or semantic construction. However, in some cases this method may still be useful, especially for users who believe, as we do, that the purpose of any visualization is not to replace human reading but to send us back to the poem. An unexpected by-product of our tool, it is already provoking rich readings, which show that shaping our queries about sound so that the machine can answer with metaphor helps us understand dynamics inherent to poetry, which invites readers to connect every feature to every other feature. Thus, POEMAGE furthers the larger goal of reliably identifying metaphor through computational methods. It also suggests ways forward in our overarching goal of identifying such complex features for visualisation in their own right.

These are just a few instances among many in which POEMAGE queries, by locating sonic difference in similarity, have identified places in poems that are not only sonically but metaphorically rich. Though we often believe the complexity of language may create an obstacle to the computational analysis of poetry, this argument represents a re-theorization of ways in which a tool originally meant to aid in one kind of analysis can give access to information and insight not originally predicted or even sought, not in spite of but because of poetry's linguistic complexity.

Bibliography

Abdul-Rahman, A.J, Walton, S., Bemis, K., Lein, J.G., Coles, K., Silver, D., and Chen, M. (2017). “Re-spatialization of time series plots.” Information Visualization, *Sage Journal*, forthcoming.

Abdul-Rahman, A., Lein, J.G., Coles, K., Maguire, E., M. Meyer, M., Wynne, M., Trefethen, A.E., Johnson, C., and Chen, M. (2013). Rule-Based Visual Mappings—With a Case Study on Poetry Visualization. *Computer Graphics* Forum 32, pp. 381-390.

Chaturvedi, M., Gannod, G., Mandell, L., Armstrong, H., and Hodgson, E. (2013). Myopia: A Visualization Tool in Support of Close Reading. *Digital Humanities 2012*. Hamburg, Germany. 18 July 2012 .

Coles, K. (2015) “Ghost (in the) Machine.” Keynote lecture. Australasian Association of Writers and Writing Programs Annual Conference. Melbourne. Dec. 1,.2015

Coles, K. (2013). “I Don’t Care About Data.” Panel presentation. Digging Into Data Challenge Round Three End-of-Project Conference. Glasgow, Jan. 27, 2016

Coles, K. (2015). “In Motion in the Machine.” Invited lecture. Poetry on the Move/International Poetry Studies Institute. Canberra. Sept. 2015.

Coles, K. (2016) “Show Ambiguity: Collaboration, Anxiety, and the Pleasures of Unknowing.” #Vis4DH, InfoVis 2016. Baltimore, October 24, 2016. *IEEE Transactions on Visualization and Computer Graphics* 23, (forthcoming, 2017).

Coles, K. (2014). Slippage, spillage, pillage, bliss: Close reading, uncertainty, and machines. Western Humanities Review, pages 39–65.

Coles, K. and Lein, J. (2014). Turbulence and Temporality: (Re)visualizing Poetic Time. *Things My Computer Taught Me About Poems.* MLA2014. Chicago, IL.

Coles, K. and McCurdy, N. (2016) Developing and Sustaining Collaborative Research in the Humanities. Panel Discussion. MLA2016. Austin, TX. Jan. 2016

Dickinson, E. (n.d). Facsimiles in the Emily Dickinson Archives, Amherst College Digital Collections, Amherst.College Library. Open source.

Hirsch, E.D. (1967). *Validity in Interpretation.* New Haven: Yale UP, 1967

Lein, J.G. (2015). Computers in my Classes: A Pedagogy Round-Table on Workshopping (With) the Digital. Panel Discussion. AWP2015. Minneapolis, MN. April 2015

Lein, J.G. (2015) “Digital Humanities and Dickinson’s ‘Tell’: Recounting Poetic Encounter.” New Work on Dickinson: Flash Talks. Modern Language Association. Vancouver, BC Jan. 2015

Lein, J.G. (2012) “Seeing the Sonic: Aesthetics, Poetry, and Data Visualization.” Aesthetics Reloaded. Aarhus, Denmark. Dec. 2012.

Lein, J.G. (2014) Sounding the surfaces: Computers, context, and poetic consequence. Western Humanities Review, pages 84–109.

McCurdy, N., Dykes, J., and Meyer, M. (2016). "Action Design Research and Visualization Design." Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization. ACM, 2016.

McCurdy, N., Lein, J.G., Hurtado, A. (2015). “Deep in Poetry: Improvisations in Tech, in Text, in Time.” IEEE VISAP 2-15. Chicago, IL.

McCurdy, N, J. Lein, K. Coles, and M. Meyer. (2016). Poemage: Visualizing the Sonic Topology of a Poem. *IEEE Transactions on Visualization and Computer Graphics*. 22:1, pp. 439-448.

McCurdy, N., Srikumar, V., and Meyer, M. (2015). RhymeDesign: A Tool for Analyzing Sonic Devices in Poetry. In *Proceedings of Computational Linguistics for Literature,* pp. 12-22.

Ramsay, S. (2011). Reading Machines: Toward an Algorithmic Criticism. Topics in the Digital Humanities. Ed. Susan Schreibman and Raymond C. Siemens. Chicago: U of Illinois Press.

Tracing Swerves of Influence: text reuse and the reception of Lucretius in 18th-century England

Charles Cooney
chu.cooney@gmail.com
University of Chicago, United States of America

Clovis Gladstone
clovisgladstone@gmail.com
University of Chicago, United States of America

Over the past several years, as part of a Digging into Data Grant and in conjunction with Oxford University’s e-Research Centre, the ARTFL Project at the University of Chicago has been developing a large-scale online resource that allows scholars to examine text reuse in the 18th century. We used sequence alignment algorithms to compile a database of textual repetitions found in the Gale-Cengage *Eighteenth Century Collections Online* (ECCO) corpus, which contains the 200,000 works that represent most of the printed literary and scientific output in Britain from 1700 to 1799.

At Digital Humanities 2016 in Krakow, we presented the methodology and algorithms we used to identify these reuses and showed the early results of our work (Roe et al, 2016; for earlier discussions of this project, see Roe et al, 2015 and Abdul-Rahman et al, 2016). For the upcoming

Digital Humanities conference, our presentation will have two facets. First, we will outline the technical and editorial approaches we took when building the final version of the alignment database to maximize its usability and usefulness as a scholarly resource (See the *Common Place Cultures* project site, and its page on the University of Chicago domain). Secondly, we will discuss a use case of the database in which we examined reuses and citations of the first-century BC Roman poet, Lucretius, as a means to get a broad understanding of the 18th-century reception of the *De Rerum Natura*, the philosophical poem that proposes a materialist conception of the universe.

To construct the database, we used the PhiloLine (see Horton et al, 2010) sequence aligner to identify many millions of similar passages in the often low-quality OCR of the ECCO dataset. These passages range from a handful of words to large extracts of documents. We then used a similar passage matching algorithm to identify passages that were reused many times. The resulting database allows users to track specific passages, identify citations of specific texts in later texts, and find borrowed passages in later documents of an author's oeuvre.

As the creators and users of our navigational tool, we had to decide on the nature and scope of research the database should support and then strike balances between feasibility, usability, and performance. One of our earliest concerns was to allow users to get a fairly long view of textual reception/citation. To be able to identify the true source of any given reuse, we made the editorial decision to include texts that predated the 18th century. We therefore extended our alignment detection procedure to a variety of curated datasets such as the King James Bible, Classical Latin texts, and EEBO-TCP.

Our extended alignment experiment was extremely successful: we uncovered more than 40 million text reuses across our multiple datasets. At the same time, this success raised the problem of devising a way to explore result sets of a huge scale efficiently, leading us to focus on building a navigation tool that provides filtering and sorting control to users in a precise and intuitive way. We added various UI elements to guide users in their exploration: a list of the most commonly cited authors just a click away from the input box; a faceted browser to help users narrow down search results; and a timeline view of any given text reuse (see Appendix for screenshots) Combined together, these choices greatly enhance the capabilities of our web application, making it a tool that can very easily track intellectual influence from the classical Latin era to the late 18th century.

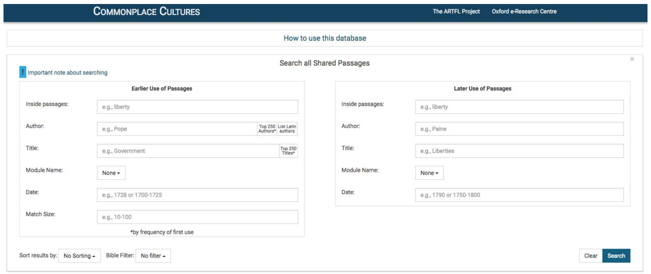
In the second part of our presentation we will discuss a use case of our database, examining citations of Lucretius's Latin text in 18th-century English texts. Our aim was to fill in gaps in current scholarship and discover the aftereffects of the resurgence in interest in Lucretius's work, the *De Rerum Natura* (*DRN*), at the middle of the 17th century in England (for example, Greenblatt (2011) gives only very cursory treatment to reception of Lucretius in the 18th century in *The Swerve*). During this so called "Epicurean revival," John Evelyn (1656) and Lucy Hutchinson (unpublished) were the first to translate, either in part or whole, the *DNR* into English. Walter Charlton published his *Physiologia Epicuro-Gassendo-Charltoniana, or, A fabrick of science natural, upon the hypothesis of atoms* in 1654. And though Thomas Hobbes never cites Lucretius directly in his *Leviathan* (1651), the Latin poet's ideas about the material nature of the universe are a distinct antecedent to Hobbes's mechanical philosophy.

Even around the time of the revival, the reception of the *DRN* was vexed. Lucretius's statements about the materiality and mortality of the soul, the role of chance in the universe, and the detachment of the gods were far too radical ever to gain wide acceptance in early modern England. In the backlash against Hobbesianism in the 1660s, Lucretius was a prime target for criticism. Cottegnies (2016) argues that the backlash against Hobbes's ideas in the 1660s also marks the end of the Epicurean moment. Theologians writing against Hobbes attacked Lucretius's atheistic materialism. And though Lucretius's stature as a poet continued to grow, praise for *DRN* was almost always tempered. His more extreme ideas were to be dismissed or ignored. In the notes to his translation of *DRN* in 1682, the first complete translation published in English, Thomas Creech argued against Lucretian and Epicurean attitudes toward the soul and divinity (Creech and Dryden, 1700; these citations refer to the text published using the PhiloLogic build of eebo). John Dryden, the preeminent arbiter of literary taste of his era, quoted Lucretius often in his plays and included translations of select passages in *Sylvae* (1685). In his preface to that collection, Dryden praised the directness of Lucretius's poetic expression, pointing out the "positive assertion of his Opinion" and his "Magisterial authority." But the subject of his poem is "naturally Crabbed" and the poet himself is "often in the wrong" (Dryden, 1685; text published online by Philologic).

Reuses and citations found in the Digging into Data database suggest that this basic framework for understanding Lucretius largely played out across the 18th century. Lucretius was at once an admired poet, a materialist attacked for not admitting divine involvement in the universe, and a philosopher who in fact had important things to say about living well. Even so, the alignment database allows us to see a handful of authors, mostly medical and scientific writers, whose views of Lucretius and his ideas veer from this basic narrative. Mainly toward the middle and latter parts of the century, some reuses suggest a less troubled acceptance of Lucretius's naturalism.

Through this presentation, we hope to show that this alignment database, through the accumulation of so many instances of citation, can facilitate a kind of large-scope reading that allows scholars to gain a nuanced sense of longer term intellectual trends. Built on a huge quantity of uncorrected OCR, the database provides scholars the specific source evidence -- and a ready means to access it -- that they might need as a starting point to pursue even deeper investigations into the thought of the 18th century.

Appendix: Screenshots



Earlier Use of Passages

Inside passages: e.g., liberty
Author: e.g., Pope
Title: e.g., Government
Module Name: None
Date: e.g., 1728 or 1700-1725
Match Size: e.g., 10-100

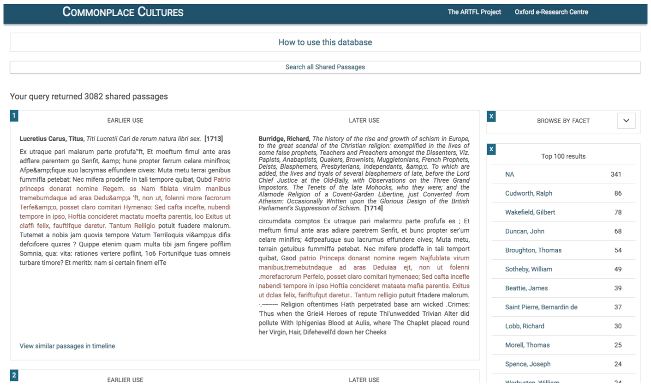
*by frequency of first use

Later Use of Passages

Inside passages: e.g., liberty
Author: e.g., Paine
Title: e.g., Liberties
Module Name: None
Date: e.g., 1790 or 1750-1800

Sort results by: No Sorting Bible Filter: No filter

Figure 1: The Commonplace Cultures search form



Your query returned 3082 shared passages

Lucretius Carus, Titus, *Titi Lucretii Cari de rerum natura libri sex.* **[1713]**

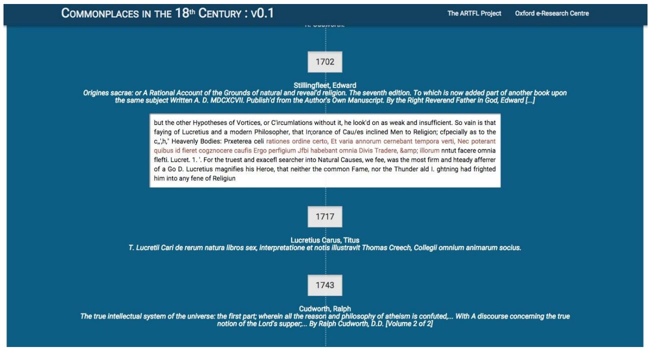
Ex utraque pari malarum parte profufa"ft, Et moeftum fimul ante aras adflare parentem go Senfit, &amp; hune propter ferrum celare minifIros; Afpe&amp;fique suo lacrymas effundere civeis: Muta metu terrai genibus fummifla petebat: Nec mifera prodeffe in tali tempore quibat, Qubd Patrio princeps donarat nomine Regem. ss Nam fiblata viruim manibus tremebumdaque ad aras Dedu&amp;a 'ft, non ut, folenni more facrorum Terfe&amp;o, posset claro comitari Hymenao: Sed cafta incefte, nubendi tempore in ipso, Hoftia concideret mactatu moefta parentis, loo Exitus ut claffi felix, fauftIfque daretur. Tantum Relligio potuit fuadere malorum. Tutemet a nobis jam quovis tempore Vatum Terriloquis vi&amp;us difis defcifcere quxres ? Quippe etenim quam multa tibi jam fingere poffIim Somnia, qua: vita: rationes vertere poflint, 1o6 Fortunifque tuas omneis turbare timore? Et meritb: nam si certain finem elTe

Burridge, Richard, *The history of the rise and growth of schism in Europe, to the great scandal of the Christian religion: exemplified in the lives of some false prophets, Teachers and Preachers amongst the Dissenters, Viz. Papists, Anabaptists, Quakers, Brownists, Muggletonians, French Prophets, Deists, Blasphemers, Presbyterians, Independants, &amp;c. To which are added, the lives and tryals of several blasphemers of late, before the Lord Chief Justice at the Old-Baily, with Observations on the Three Grand Impostors. The Tenets of the late Mohocks, who they were; and the Alamode Religion of a Covent-Garden Libertine, just Converted from Atheism: Occasionally Written upon the Glorious Design of the British Parliament's Suppression of Schism.* **[1714]**

circumdata comptos Ex utraque pari malarmru parte profufa es ; Et meftum fimul ante aras adiare paretrem Senfit, et bunc propter ser'um celare minifirs; 4dfpeafuque suo lacrumus effundere cives; Muta metu, terrain getuibus fummiffa petebat. Nec mifere prodeffe in tali temport quibat, Gsod patrio Princeps donarat nomine regem Najfublata virum manibus,tremebutndaque ad aras Deduiaa ejt, non ut folenni .morefacrorum Perfelo, posset claro comitari hymenaeo; Sed cafta incefle nabendi tempore in ipso Hoftia concideret mataata mafia parentis. Exitus ut dclas felix, farìftufqut daretur.. Tantum relligio putuit frtadere malorum. ·.——— Religion oftentimes Hath perpetrated base arn wicked .Crimes: 'Thus when the Griei4 Heroes of repute Thi'unwedded Trivian Alter did pollute With Iphigenias Blood at Aulis, where The Chaplet placed round her Virgin, Hair, Difehevell'd down her Cheeks

Top 100 results	
NA	341
Cudworth, Ralph	86
Wakefield, Gilbert	78
Duncan, John	68
Broughton, Thomas	54
Sotheby, William	49
Beattie, James	39
Saint Pierre, Bernardin de	37
Lobb, Richard	30
Morell, Thomas	25
Spence, Joseph	24
Warburton, William	24

Figure 2: Search results filtered by facet



1702

Stillingfleet, Edward
Origines sacrae: or A Rational Account of the Grounds of natural and reveal'd religion. The seventh edition. To which is now added part of another book upon the same subject Written A. D. MDCXCVII. Publish'd from the Author's Own Manuscript. By the Right Reverend Father in God, Edward [...]

but the other Hypotheses of Vortices, or C'ircumlations without it, he look'd on as weak and insufficient. So vain is that faying of Lucretius and a modern Philosopher, that In;orance of Cau/es inclined Men to Religion; cfpecially as to the c,,',h," Heavenly Bodies: Prxeterea celi rationes ordine certo, Et varia annorum cernebant tempora verti, Nec poterant quibus id fieret cogznocere caufis Ergo perfigium Jfbi habebant omnia Divis Tradere, &amp; illorum nntut facere omnia flefti. Lucret. 1. '. For the truest and exacefl searcher into Natural Causes, we fee, was the most firm and hteady afferrer of a Go D. Lucretius magnifies his Heroe, that neither the common Fame, nor the Thunder ald I. ghtning had frighted him into any fene of Religiun

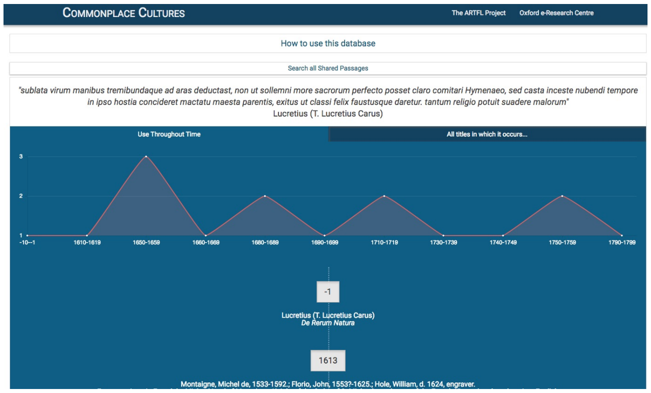
1717

Lucretius Carus, Titus
T. Lucretii Cari de rerum natura libros sex, interpretatione et notis illustravit Thomas Creech, Collegii omnium animarum socius.

1743

Cudworth, Ralph
The true intellectual system of the universe: the first part; wherein all the reason and philosophy of atheism is confuted,... With A discourse concerning the true notion of the Lord's supper;... By Ralph Cudworth, D.D. [Volume 2 of 2]

Figure 3: Search results by year



"sublata virum manibus tremibundaque ad aras deductast, non ut sollemni more sacrorum perfecto posset claro comitari Hymenaeo, sed casta inceste nubendi tempore in ipso hostia concideret mactatu maesta parentis, exitus ut classi felix faustusque daretur. tantum religio potuit suadere malorum"
Lucretius (T. Lucretius Carus)

-1

Lucretius (T. Lucretius Carus)
De Rerum Natura

1613

Montaigne, Michel de, 1533-1592.; Florio, John, 1553?-1625.; Hole, William, d. 1624, engraver.

Figure 4: Search results with timeline

Bibliography

Abdul-Rahman, A., Roe, G., Olsen, M., Gladstone, C., Whaling, R., Cronk, N., Morrissey, R. and Chen, M. (2016). "Constructive Visual Analytics for Text Similarity Detection." Computer Graphics Forum

Cottegnies, L. (2016) "Michel de Marolles's 1650 Translation", pp. 161-189 in Norbrook, Harrison, and Hardie eds, *Lucretius and the Early Modern* (Oxford: Oxford University Press).

Creech, T. and Dryden, J. (1700) *Lucretius his six books of epicurean philosophy and Manilius his five books containing a system of the ancient astronomy and astrology together with The philosophy of the Stoicks / both translated into English verse with notes by Mr. Tho. Creech; To which is added the several parts of Lucretius, English'd by Mr. Dryden*. London and Westminster, London. Published online by Philologic. (EEBO-TCP; phase 1, no. A49437) Transcribed from Early English Books Online; image set 45711.

Dryden, J. (1685) *Sylvæ, or, The second part of Poetical miscellanies* London, Tonson. Published online by Philologic. (EEBO-TCP; phase 1, no. A36697) Transcribed from: Early English Books Online; image set 58020

Edelstein, D., Morrissey, R., and Roe, G. (2013) "To Quote or not to Quote: Citation Strategies in the Encyclopédie", *Journal of the History of Ideas* Vol. 74, No. 2: 213-236.

Greenblatt, S. (2011) The Swerve (New York: W.W. Norton).

Horton, R., Olsen, M., and Roe, G. (2010) "Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections," *Digital Studies / Le Champ numérique* Volume 2, Number 1

Roe, G., Abdul-Rahman, A., Chen, M., Morrissey, R., Olsen, M. (2015). "Visualizing Text Alignments: Image Processing Techniques for Locating 18th-Century Commonplaces." *Digital Humanities 2015,* Sydney, Australia, July 1, 2015.

Roe, G., Gladstone, C., Morrissey, R., and Olsen, M. (2016) "Digging into ECCO: Identifying Commonplaces and Other Forms of Text Reuse at Scale." *Digital Humanities 2016,* Krakow, July 13, 2016.

What News is New?: Ads, Extras, and Viral Texts on the Nineteenth-Century Newspaper Page

Ryan Cordell
rccordell@gmail.com
Northeastern University, United States of America

David Smith
dasmiq@gmail.com
Northeastern University, United States of America

Newspapers became a truly mass medium in the nineteenth century due to the steam press, the post, the telegraph, and, especially in the U.S., mass political parties. Many scholars working on newspapers and other nineteenth-century media have noted the high level of "exchange" between different publications and the importance of understanding what was reprinted for understanding what readers and editors of the period valued (see, for instance, McGill, 2003; or Garvey, 2012).

In the first stages of the Viral Texts Project, we developed efficient clustering methods based on statistical language modeling and alignment to identify reprinted texts in digital archives of newspapers and magazines, beginning with the Chronicling America corpus and expanding to include papers from the UK, Australia, and Europe. Approaching the newspaper corpus through the lens of text reuse, Viral Texts has led to substantial insights into low-level problems of text reuse analysis in errorful OCR archives, higher-level network analysis of cultural circulation (Smith et al, 2015), the informational mode of nineteenth-century newspaper reprinting, the network effects of authorship in the nineteenth-century newspaper medium (Cordell, 2015), the circulation of popular "fugitive poetry" during the period (Cordell and Mullen, forthcoming 2017), the bibliographic implications of errorful OCR (Cordell, forthcoming 2017), among other things (project publications and press are available on the project site, and the project data sets are available on Github).

The computational methods of the Viral Texts Project produce a database view of the textual field. We read nineteenth-century newspaper snippets as "clusters" of reprints in a spreadsheet or database. We do not read them in the contexts of their original publications, but as disambiguated segments of text excerpted, aggregated, and listed. Such a database view is not *substantially* distinct from much older bibliographic views. The "clusters" of reprints generated by this algorithmic approach are still, essentially, enumerative bibliographies of textual snippets that circulated in nineteenth-century newspapers. Organized in a database, their appearance and presentation echo the conventions of printed bibliographies that list, for example, all known witnesses of a particular author's works.

Considering a text's bibliography allows us to speak of it in terms of circulation, audience, and influence, but the meanings of those corpus-scale phenomena are not evenly distributed among the witnesses that comprise its bibliography, when considered individually at the codex scale. Our current work seeks to map what we know of reprinting at a systemic level back onto the newspaper page, taking up the challenge in Matthew Philpott's recently articulated "understanding of the conceptual dimension of a periodical as a statistically self-similar, fractal form across all levels of scaling, from the periodical as a whole in its full publication run, to the year or annual volume, and down to the individual issue or number." (Philpotts, 2015) Likewise, what we are learning about disambiguated reprinted texts can help us understand the generic conventions and material operations of particular newspapers from our corpus over time, and to compare such features among papers.

Since developing our methods to detect reprints, project Ph.D. student Jonathan Fitzgerald has developed classification methods for sorting the millions of resultant clusters into meaningful categories: separating poetry from prose, for instance, or fiction from advertising testimonials and hard news (for discussion of some of this early classification work, see Fitzgerald, 2016). Building on Fitzgerald's classification work, we can now ask what generic trends we can spot within particular newspapers, or across the corpus. Some genres prove quite amenable to automatic classification, with success rates well over 90% for classifying advertisements and news and 84% for classifying poetry. Nineteenth century scholars and book historians, for instance, would expect to see poems in particular corners of the newspaper: the top left corner of page one in some papers, or the top left corner of page 4 in others. Drawing on tens of thousands of automatically identified poems, we can test those ideas more broadly, asking just how consistently the *Lewisburg Chronicle, and the West Branch Farmer* reproduced poems in the page one poetry corner, or just what percentage of papers chose one or the other of these two conventional placements. Perhaps more interestingly, we look for outliers in the data: newspapers that printed poetry in ways that defied generic norms, which may prove to be particularly interesting for periodicals and book historians.

Another structural aspect of periodicals addressed by this approach could be summed up with the question, "What news is new?" In short, while we know newspapers relied on reprinting and other kinds of textual recycling to fill column inches with limited staff, we can now compare habits of reuse within papers and among papers, asking for instance which newspapers relied more or less on reprinting—which were primarily producers or consumers of content—and how reprinted content populates the pages of particular newspapers. Again, we ask what patterns of new vs. reprinted content we can map across issues of particular papers, or how those patterns compare among different papers. Finally, we investigate how reprinted material migrates within particular newspapers over time; as stories age, can we trace them moving from page one through the latter pages of subsequent issues?

At an even finer level, we can ask how much total content was kept set up in type from one newspaper issue to the next. In the early and middle nineteenth century, many newspaper advertisements ran continuously for weeks or months at a time. Some advertisements even had short codes printed at the end for start date and duration, so that compositors could check whether an ad should be kept. In addition, news stories might be only slightly updated from day to day. By tracing this reuse among consecutive issues of the same newspaper, we can better estimate how much work by editors and compositors went into putting together the paper. We can also see how these practices changed over the course of the century as wire services and national advertising campaigns increased the freshness of newspaper content (figure 1). By illuminating precisely what text was kept in standing type, both in particular newspapers and comparatively across the corpus, we generate new insight into the workings of the nineteenth-century printing office, as well as the priorities and work habits of editors and other newspaper employees. Given the new internationalism of our corpus, we can also compare such practices across national and linguistic boundaries.

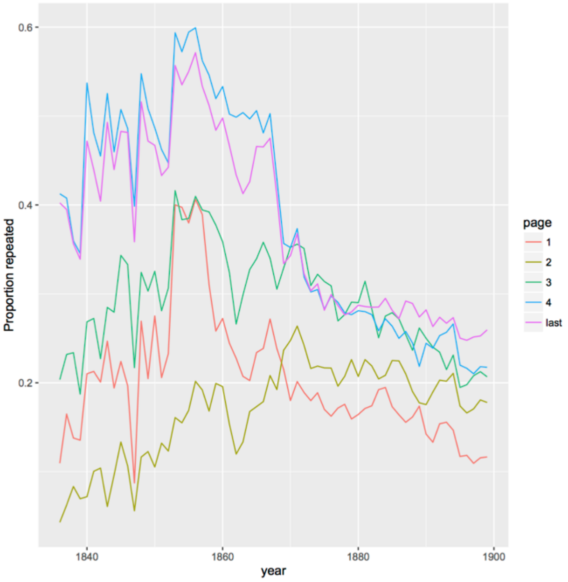


Figure 1: Average proportion of text on a page repeated from the previous issue of the same newspaper in the Chronicling America dataset, 1835–1900. Separate lines show reprint proportions on page 1, 2, 3, and 4. Through the early part of this historical period, most newspaper issues had four pages. A separate line shows reprint proportion on the last page, whatever it is. Repeated content is measured by globally aligning each page with all pages of previous issues in the last seven days.

For this paper, we build on the alignment and clustering methods developed in the Viral Texts project in two ways. First, by applying text categorization at the cluster level, we are able to make more robust inferences about genre than we would by scanning an individual newspaper page. Since each cluster of reprints preserves the location on the original page of each witness, we can then easily map these inferences back onto each issue. Second, we implement new, efficient global alignment algorithms to trace repeated ads, boilerplate, and news updates across successive issues of all newspapers in the corpus. We can prune the search space of this alignment algorithm more aggressively since texts kept in set type will retain the same line breaks, spelling errors, etc., as their previous printings. For robustness in the face of OCR errors, we align each issue to one or two weeks worth of previous issues. While global, linear (monotonic finite-state) alignment gives us promising initial results (see figure 1), we are also experimenting with greedy top-down inference that allows block moves of passages. Finally, we note that this procedure for detecting repeated passages across consecutive issues gives us yet more evidence about the boundaries of stories, which might not be typographically marked by, e.g., headline fonts.

These are just a few examples of how we are modeling the structure of newspaper layout and production. Such modeling allows us to test ideas about newspaper materiality advanced by scholars working with particular publications and ask new questions about both trends and outliers in the newspaper system of the nineteenth century.

Bibliography

Cordell, R. (2015) "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History* 27.3 (August 2015), pre-print available at http://ryancordell.org/research/reprinting-circulation-and-the-network-author-in-antebellum-newspapers/.

Cordell, R. (forthcoming, 2017) "'Q i-jtb the Raven': Taking Dirty OCR Seriously,", forthcoming in *Book History*.

Cordell, R., and Mullen, A. (forthcoming, 2017) "'Fugitive Verses': The Circulation of Poems in Nineteenth-Century American Newspapers," forthcoming in *American Periodicals* 27.1 (Spring 2017), preprint available at http://viraltexts.org/2016/04/08/fugitive-verses/.

Fitzgerald, J. D. (2016). "Computationally Classifying the Vignette Between Fiction and News". *Jonathan D. Fitzgerald.* Blog post. 10 October 2016. Available at http://jonathandfitzgerald.com/blog/2016/10/10/the-viral-vignette.html

Garvey, E. G. (2012) *Writing with Scissors: American Scrapbooks from the Civil War to the Harlem Renaissance* (Oxford: Oxford University Press).

Philpotts, M. (2015) "Dimension: Fractal Forms and Periodical Texture," *Victorian Periodicals Review* 48.3 (Fall 2015): 413.

McGill, M. (2003) *American Literature and the Culture of Reprinting, 1834-1853* (Philadelphia: University of Pennsylvania Press).

Smith, D., Mullen, A., and Cordell, R. (2015) "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers," *American Literary History* 27.3 (August 2015), preprint available at http://viraltexts.org/2015/05/22/computational-methods-for-uncovering-reprinted-texts-in-antebellum-newspapers/.

Digital Humanities from Scratch: A Pedagogy-Driven Investigation of an In-Copyright Corpus

Brian Croxall
brian.croxall@brown.edu
Brown University, United States of America

Following the publication of Franco Moretti's *Graphs, Maps, Trees*, scholars looking to apply digital humanities methods to literature have increasingly been drawn to "distant reading." The influence of distant reading in digital humanities is apparent not only in the work it has inspired (see, among others, Cordell and Smith; Elson, Dames, and McKeown; Jockers; Long and So; Rhody; and Underwood) but also for its regular inclusion as a method in courses introducing DH. "Teaching digital humanities," it turns out, often means "teaching distant reading."

Teaching students the techniques of distant reading can be challenging as it depends on re-framing the familiar object of study. But another difficulty altogether is that this approach depends on a digitized corpus; and such a corpus, in turn, depends on someone, somewhere doing the difficult labor of digitization. One might ask, then: if "teaching digital humanities" means "teaching distant reading," shouldn't it also mean "teaching digitization"?

In this paper, I will discuss a collaborative, multi-year assignment that I conducted in two of my "Introduction to Digital Humanities" courses at Emory University: the digitization and analysis of the complete works of Ernest Hemingway (Croxall). With the goal of teaching my students not only how to do distant reading but also about the intense labor that goes into corpus preparation, we digitized the whole of Hemingway's work in just two weeks. Working from newly purchased copies of the texts, the students and I rapidly scanned hundreds of pages, performed and corrected optical character recognition, and assembled a corpus—with each of us spending no more than 4 hours on the task. Our from-scratch corpus was composed expressly so we could draw important distinctions among Hemingway's works: individual works vs the whole collection; fiction vs non-fiction; and works published before while Hemingway was alive vs those that appeared after his death in 1961. I will detail what we learned from rapid digitization and how those lessons affected the second iteration of the assignment.

After preparing the corpus, students worked in groups to analyze the many works of Hemingway that they had not had time to read. Making use of Voyant Tools, they identified themes in the corpus and charted patterns that could never have been observed through regular, close reading methods. For example, the class confirmed that while Hemingway insists on writing about "men," the women to whom they are attached are inevitably just "girls." In an attempt to chart the patterns of Hemingway's diction, another group of students investigated the terms he uses to introduce dialogue. Unsurprisingly, the students discovered that "said" is by far the most frequent such term across the entire corpus. What was more surprising, however, was to observe that in late and posthumous writings, the frequency of "said" suddenly drops by 50%. In short, by building our own corpus from scratch, the students were able to conduct original research, something that is relatively rare for many undergraduates in humanities programs.

Building our collection of texts from scratch had two critical advantages. First, we were able to create a small, relatively clean corpus whose provenance we knew. This provided a sense of confidence in the data as we began to distant read. Furthermore, while our analysis of Hemingway's works was "distant" compared to traditional close reading of a single novel or story, it was not nearly as distant as projects that deal with several thousand texts. We became engaged, in short, in close-distant reading. Second, digitizing the texts ourselves allowed us to skirt a problem that frequently plagues distant reading texts from the twentieth century: copyright. As an educational endeavor focused on teaching the students how to prepare their research materials, this guerilla digitization project fell under the regime of fair use in the United States.

To close, I will discuss how students at Brown University and I have taken further steps with the Hemingway corpus and with their digital humanities education as we have used it as a means to explore the methods and utility of topic modeling. Topic modeling is frequently deployed to come to terms with large and unwieldy corpora (see Jockers; Nelson; Nelson, Mimno, and Brown; Underwood and Goldstone). But working with a small, relatively clean corpus that is created from scratch allows students to better understand what takes place via unsupervised machine learning. At the same time, topic modeling allows us to ask in a new way some of the same questions that my former students had already uncovered: how does Hemingway's dialog differ from his prose? how different are the topics in Hemingway's fiction from those of his non-fiction? to what degree does his late—or even posthumous—work differ from what he wrote three decades earlier?

In the end, the process of modeling Hemingway becomes a means by which we can model all of digital humanities—both analysis and corpus creation—in a student-focused environment (see also Brier; Croxall and Singer; Harris; Hirsch; Jewell and Lorang; and Swafford). By doing digital humanities from scratch, students can be engaged in original research and see for themselves, from start to finish, how digital humanities gets done.

Bibliography

Brier, S. (2012). "Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities." In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minnesota University Press, pp. 350-367.

Cordell, R. and Smith, D. A. (2017). *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines*. http://viraltexts.org/ (accessed 7 April 2017).

Croxall, B. (2015). "How to NOT Read Hemingway." *Intro to DH*. http://www.briancroxall.net/s15dh/assignments/how-to-not-read-hemingway/ (accessed 7 April 2017).

Croxall, B. and Singer, K. (2013). "The Future of Undergraduate Digital Humanities." Digital Humanities 2013, Lincoln, NE, July 2013.

Elson, D. K., Dames, N. and McKeown, K. R. (2010). "Extracting Social Networks from Literary Fiction." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguisticsi,* Uppsala, Sweden.

http://www.cs.columbia.edu/~delson/pubs/ACL2010-ElsonDamesMcKeown.pdf (accessed 7 April 2017).

Goldstone, A. and Underwood, T. (2014). "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45.3: 359-384.

Harris, K. D. (2011). "Pedagogy & Play: Revising Learning through Digital Humanities." Digital Humanities 2011, Stanford, CA, June 2011.

Hirsch, B. D. (2012). *Digital Humanities Pedagogy: Practices, Principles and Politics*. Open Book Publishers.

Jewell, A. and Lorang, E. (2016). "Teaching Digital Humanities Through a Community-Engaged, Team-Based Pedagogy." Digital Humanities 2016, Kraków, Poland, July 2016.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana Champaign, IL: University of Illinois Press.

Long, H. and So, R. J. (2016). "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning." *Critical Inquiry* 42.2: 235-267.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Moretti, F. (2007). *Graphs, Maps, Trees*. London: Verso.

Nelson, R. K. (2011). *Mining the Dispatch*. http://dsl.richmond.edu/dispatch/ (accessed 7 April 2017).

Nelson, R. K., Mimno, D. and Brown, T. (2012) "Topic Modeling the Past." Digital Humanities 2012, Hamburg, Germany, July 2012.

Rhody, L. M. (2013). "Revising Ekphrasis: Methods and Models." *The Association for Computers and the Humanities.* http://ach.org/2013/12/30/revising-ekphrasis-methods-and-models/ (accessed 7 April 2017).

Sinclair, S. and Rockwell, G. (2017). *Voyant Tools*. http://voyant-tools.org/ (accessed 7 April 2017).

Swafford, J. E. (2016). "Read, Play, Build: Teaching Sherlock Holmes through Digital Humanities." Digital Humanities 2016, Kraków, Poland, July 2016.

Underwood, T. (2013). *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies*. Stanford: Stanford University Press.

A World of Difference: Myths and misconceptions about the TEI

James C. Cummings
james.cummings@it.ox.ac.uk
University of Oxford, United Kingdom

Introduction

The Guidelines of the Text Encoding Initiative are generally recognised in the digital humanities as important and foundational standards for many types of research in the field. The Guidelines of the TEI are generalistic, seeking to enable the largest possible user base encoding digital texts for a wide range of purposes. Working on many TEI-based projects, teaching TEI workshops, and advising researchers on data modelling needs, I have encountered many misunderstandings about the TEI. Indeed, one keynote lecture (not at DH) once told me that "the problem with the TEI is it has too many tags and there is no way to change it". Inspired by myths like this, this paper will detail and expose common misconceptions about the TEI -- all of which have been espoused to me at some point -- but will concentrate on the more technical myths in a hope to increase knowledge about the TEI while dispelling some misconceptions along the way. Some of those to be investigated include:

"The TEI is too big (or complicated)"

While there is some truth to this -- the TEI Guidelines are numerous, consisting of around 565 elements -- no single project needs them all. Indeed, the TEI has mechanisms for customisation and recommends doing so to any project. The Guidelines themselves are modular and not all chapters will be appropriate or necessary to read for all projects.

"There is no way to change the TEI"

Although I have heard even well-respected keynote lecturers (not at DH) espouse this belief, it is patently and demonstrably false. This myth arises from unfamiliarity with the fact that the TEI is a framework entirely based on the concepts of adaptability and modification. Not only does the TEI have a sophisticated literate programming methodology to create meta-schemas which subset, constrain, and extend the vocabulary for any individual encoding project, but it also provides a variety of tools to enable users to do so.

"The TEI is too small (or doesn't have <my:SpecialElement>)"

While seemingly the opposite of #1, a frequent complaint made by those unfamiliar with the customization mechanisms of the TEI is that it does not have the special element needed for a particuar encoding project. There is, naturally, a reluctance to add new elements to one's customization -- and getting more generalized solutions into the TEI Guidelines themselves is indeed a better solution -- however, many new elements are added to the Guidelines through community development across disciplines. Any user is free to add <my:SpecialElement> but generally it is a better idea to get a number of individuals or a special interest group to agree a more detailed proposal.

"The TEI is XML (and XML is broken or dead)"

This idea is usually espoused by those who want to support some other, newer, format. Leaving aside the need some feel to denigrate one format in order to support another, XML is a widely supported format which will be with us for many years to come. However, TEI is **not** XML -- it is currently serialized as such, but previously it has been serialized as SGML, and in the future it may be expressed in another format(s). While there is currently no other widely adopted format which meets the many and varied needs of the TEI's central format, this does not mean that the TEI cannot be used with many other formats (as input, output, integrated with it).

"XML (and thus TEI) can't handle overlapping hierarchies"

Many people have discussed their concerns of overlapping hierarchies in XML, and while it is true that there are limitations in expressing multiple hierarchies in XML, it also has solutions built into it, such as empty elements to represent one or more alternative hierarchies. Primarily, this misunderstanding is also based on the assumption that all markup is embedded markup. The TEI Guidelines include a chapter on representing non-hierarchical structures, and the TEI framework has many features for representing fragmented element structures, out-of-line and stand-off markup, and the association of additional annotation through URI-based pointing. In addition, many DH text encoding projects only require two hierarchies (e.g. intellectual vs physical representations) and the TEI provides transformation solutions to alternate between these.

"You can't do stand-off markup in XML (or TEI)"

This myth shows a misunderstanding of both XML and the TEI. The former is a language for markup vocabularies and puts no restriction on whether that markup is embedded, out-of-line, or entirely stand-off. The TEI Guidelines provide a number of solutions entirely geared to stand-off markup, and its community is working towards introducing more features in this area. The combination of fine-grained markup, URI-based pointing and/or XPointer schemes, and descriptive markup designed to function this way, means that stand-off markup is supported in the TEI.

"You can't get from TEI to $myPreferredFormat"

One of the benefits of XML is that it is easily processable to other formats. The TEI Consortium provides around 40 conversions to/from other formats, including, for example: bibtex, cocoa, csv, docbook, docx, dtd, epub, html(5), xsl-fo, json, InDesign, latex, markdown, mediawiki, nlm, odd, pdf, rdf, relaxng, slides, txt, wordpress, xlsx, xsd, and many more. There exist RESTful web services like *OxGarage* which can provide a pipeline for these and other conversions.

"There are no tools that understand the TEI"

This is false -- thousands of TEI projects have created many tools which process, mine, convert, and visualize TEI data. While the TEI Wiki lists some of these, one of the problems is that projects do not necessarily advertise and openly share their tools. Much of the software developed by projects is also bespoke and specific -- they are not necessarily generalisable to other projects' needs. There are also many sophisticated encoding activities (such as stand-off markup) for which there are few general tools, since these are usually implemented in project-specific methods.

"If you create a TEI-based digital edition you must learn other $tech"

While historically it has been the case that to create TEI-based digital editions one must learn, or employ those who know, various technologies, this is increasingly less of an issue. Out of the box software like *eXist-db's TEI Publisher* and *TEI Boilerplate* mean researchers are able to publish digital editions for themselves. Moreover, the TEI has introduced implementation-agnostic methods for documentation of intended processing models in a TEI customization. This can then be used to generate project-specific code based on changes to the customization, as in the case of the eXist-db implementation of the TEI processing model. This new aspect of the TEI enables developers to write more generalized software which relies on the TEI ODD customization file for information on the processing model.

"TEI is only for Anglo/Western works"

There is much about the TEI Guidelines that is based in Anglo and Western European textual traditions, but the Guidelines also make an effort to enable use in other languages and cultures. The definitions and glosses of elements (etc.) can be viewed in a number of languages (English, German, Spanish, French, Italian, Japanese, Korean, Chinese). There is an entire internationalization framework built into the TEI Guidelines and the TEI Customization language, which means that the schemas can routinely display these internationalized definitions in editors and those creating customisations can have definitions, examples, and attribute value descriptions, in any Unicode-expressable language.

"Interoperability is impossible with the TEI"

Interoperability is a good and laudable goal, but the potential richness of TEI encoding for research and analysis purposes should not be sacrificed for this (depending on the point of the initial encoding project). While interoperability does suffer in a framework that is customizable and extendable (which are necessary for such a generalized system), it is certainly possible. Usually it is a process of crosswalks or some scripted transformation to a lowest common denominator that involves someone knowing both resources. The creation of sub-communities (such as the TEI subset EpiDoc), which agree encoding standards that are tighter than the necessarily general and flexible TEI, can improve this significantly.

"The TEI is only for digital edition(s)"

The TEI may be used for many forms of output, for example camera-ready copy. The primary mistake here is to assume a one-to-one relationship between TEI encoded files and a single particular output. If significant encoding has taken place, a wide variety of outputs are possible. If the format is used to its full potential, many aspects of an edition can be created, as well as supplementary files, indices, introductory material, interactive data visualizations, and more. The use of the TEI can also be used outside of edition-building, for the creation of linguistic corpora, digital facsimiles, and other resources.

Summary

While these are only some of the myths surrounding the TEI, discussing these will be beneficial to the DH audience, and will hopefully lead potential TEI users to question other "received wisdom" about the Guidelines.

A Common Conceptual Model for the Study of Poetry in the Digital Humanities

Mariana Curado Malta
mariana@iscap.ipp.pt
Instituto Politécnico do Porto, Portugal

Elena González-Blanco
egonzalezblanco@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Clara Martínez Cantón
cimartinez@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Gimena Del Rio
gdelrio.riande@gmail.com
CONICET, Argentina

Many of the DH approaches to poetry have focused on its metrical aspect (Birnbaum & Thorsen, 2015a, 2015b; Pue, Teal & Brown, 2015; González-Blanco, Martínez Cantón & Rio Riande, 2015). Following Chatman & Levin (1967, p. 142) we could say that meter is a "systematic literary convention whereby certain aspects of the phonology are organized for aesthetic purposes." In this sense, versification is an abstraction of linguistic phenomena in which words (in their formal and semantical aspect) relate to rhythm and rhyme for artistic purposes. Although many theories about versification and metrics have been developed for different languages and traditions, our work is interested in a structural and formal approach that looks at poetry into discrete units, categories, and their relationships. Thus, we are involved in analysing how metrical repertoires in digital form model those structures.

A digital repertoire of poetry metrics is a catalogue that gives account of the metrical and rhythmical schemes of either a poetical tradition, a period or school, gathering a corpus of poems that are defined and classified by their main characteristics. This kind of repertoires may sometimes contain the text of the poem and information related to authors, manuscripts, editions, music, and other features, all of them related to the poems. In the beginning, repertoires were printed books in which we could find information listed in a way similar to an address book. The digital era has changed the way in which information is displayed allowing the user to perform complex and multiple searches. In all these cases there is an ontological leap when the data is put in digital format.

There are a vast number of European digital metrical repertoires available online in open access – e.g. French lyrical collections (Nouveau Naetebus), Italian (BedT), Hungarian (RPHA), Medieval Latin (Corpus Rhythmorum Musicum, Annalecta Hymnica Digitalia), Classical Latin (Pedecerto), Galician-portuguese (Oxford Cantigas de Santa María, MedDB2), Castilian (ReMetCa), Dutch (Dutch Song Database), Occitan (BedT, Poèsie Neotroubadouresque, The last song of the Troubadours), Catalan (Repertori d'obres en vers), Skaldic (Skaldic Project), or German (Lyrik des Minnesänger), among many others. Each one of these metrical repertoires was developed in a specific technology and stores data in its own database. This data is locked in the information silos of each repertoire, not available freely to be compared and used by intelligent machines that could infer over the data. The lack of interoperability between the different digital repertoires dealing with poetry across different languages, literatures and traditions is a problem that needs to be addressed (González-Blanco & Seláf, 2014; González-Blanco, Rio Riande, Martínez Cantón & Martos Pérez, 2014; González-Blanco, E. G., Rio Riande, G.del & Martínez Cantón, C., 2016a). There is an interoperability problem since the technological solutions used for building the database of each digital repertoire employ a different data model.

POSTDATA is a project funded by a Starting Grant of the European Research Council whose main goal is to study European poetry in a comparative and wide context. Its starting point is the analysis of digital metrical repertoires as digital objects that have been developed to expose European poetry on the Web of Documents in open access. POSTDATA is an idea that resulted from previous work developed by a network of partners that own digital metrical repertoires and that have been working together for some years.

The POSTDATA project focuses mainly on the possibilities of Linked Open Data (LOD) technologies in order to publish the data available on information silos in the Web of Data as LOD. A side effect of this project is that not only this network will be able to publish its own data in LOD but also any other entity that owns data related to poetic metrics. Once the data is structured the same way and published in the Web of Data, anyone will be able to deploy new software that uses this data. It will be possible to compare, extrapolate and create new views of the data, opening doors to new knowledge in i) literary research communities since POSTDATA results will enable comparisons of poetic traditions and the creation of new repertoires, and ii) literature learning environments for e.g. the development of editorial projects following the common conceptual model.

LOD technologies are very powerful since one can implement intelligent agents to infer over the data (W3C, 2010). Another advantage of LOD over other paradigms is the infinite possibilities of linking data, allowing agents to rely also on other sources of information to build knowledge.

The first step of POSTDATA is the development of a Metadata Application Profile (MAP) for the community of practice of Digital Humanities that deal with poetry. An MAP is a semantic model, a construct that enhances interoperability (Nilsson, Baker, & Johnston, 2008).

The development of an MAP is a crucial task for any community of practice. This development should be structured, and has to integrate –since the early phases of development– elements of representative members of the community. Commonly the organizations of a community of practice differ in organization-type, location, culture and in the language they speak. In poetry studies, all these are key factors, since metrics and their features are based largely on the nature of the languages. In addition, in the case of the Digital Humanities Poetry metrics community, the way metrics has been conceptualised in the different traditions is however diverse, as there are multiple ways of encoding and understanding metrical systems. Finding a common ground of understanding in such an environment becomes a huge challenge. These digital metrical repertoires are the key to build an inclusive MAP for poetry, since they reflect different approaches of conceptualising poetry and their formal features.

POSTDATA is using Me4MAP, a method for the development of metadata application profiles (see Curado Malta & Baptista 2013a, 2013b), to develop the referred MAP. According to Me4MAP, the first activities in such a development are "S1 – Developing the Functional Requirements" and "S2 – Developing the Domain Model," these two activities result, respectively, in the Functional Requirements deliverable and in the Domain Model deliverable. S1 activity feeds S2 activity.

A Domain Model is a conceptual model that is used to explicit the concepts that exist in a certain universe of discourse. The concepts have properties to show how they are defined and related.

As described in Curado Malta, Centenera & Gonzalez-Blanco (2017), the work-team is developing the Domain Model at the same time S1 activities occur, slightly changing the order of S1 and S2. This change is possible since Me4MAP is also under evaluation and testing in a Design Science Research methodological approach –for more information about this process see Curado Malta (2014). This Me4MAP testing is leaded by one of the authors.

The goal of this paper is to present a preliminary version of the Domain Model of the MAP for Poetry – Domain Model V0.1. The following paragraphs give details on how this version of the Domain Model was developed.

The Domain model V0.1 was developed based on the analysis of eleven data models of databases of metrical repertoires (see Table 1) using a reverse engineering technique –for more details about this technique see Curado Malta et al. (2017).

Project	URL
Base de Datos da Lírica profana galego-portuguesa (MedDB)	https://www.cirp.gal/meddb
Cantigas de Santa Maria	http://csm.mml.ox.ac.uk/
Corpus of Spanish Golden-Age Sonnets	https://github.com/bncolorado/CorpusSonetosSigloDeOro
Corpus Rhythmorum Musicum	http://www.corimu.unisi.it
Kalevala	http://dbgw.finlit.fi/skvr/
Lyrik des hohen Mittelalters	http://www.lhm-online.de
Métrique en Ligne	http://www.crisco.unicaen.fr/verlaine/
Repertorio Métrico Digital de la Poesía Medieval Castellana (ReMetCa)	http://www.remetca.uned.es
Répertoire métrique de la poésie lyrique occitane des troubadours à leurs héritiers (xiiie-xve siècles)	Local Database
Reperrtoire de la Poésie Hangroise Ancienne (RPHA)	http://rpha.elte.hu/
Versologie	http://metro.ucl.cas.cz/kveta

Table 1. Repertoires which databases were used as basis for Domain Model V0.1

These repertoires have different technologies and paradigms: ten are in the Web of Documents and one is deployed locally. The technologies in which the databases are implemented are: 1) MySQL databases, 2) XML databases and files, 3) Perl scripts and, 4) Worksheet files.

The Domain Model of the MAP is represented as an Unified Modeling Language (UML) class diagram that expresses POSTDATA proposal for a common conceptual model for the European Poetry. The Domain Model V0.1 is available online (permanent link). The technique used to represent the domain model is UML since it is a standard technique (ISO/IEC 19505-1 and 19505-2) to model businesses and processes very well known in the software engineering community –see Rumbaugh, Jacobson and Booch (2004).

This first version will be tested in a Workshop held in March where partners of the project will follow a hands-on session to test the Domain Model V0.1 using a resource of its own database. POSTDATA will give to each partner a template testing sheet and information about the mapping between the database in question and the Domain Model. The conclusions of the testing and of other workshop discussions will feed the process of development of the Domain Model. The work-team programs to deliver a version 0.2 out of this worshop conclusions. This process of development is highly iterative; it will be fed by the following activities:

- development of S1 activities (as already referred before);
- Analysis of the results of a survey to final users: POSTDATA wants to know the needs of final users of repertoires in order to understand their problems and how they can be solved. Part of the survey is created based on the existent models. The survey also has open questions for users to fill in freely.
- Analysis of other poetic repertoires: there are still at least eleven databases' data models to be analysed. POSTDATA is still looking for more poetic repertoires in order to have a wider representation of other languages and traditions (plewease see the geographical view of the repertoires) Analysis of the information of two case studies: two researchers are building a digital repertoire at the same time the MAP is being developed; they will inform the workteam about their needs.

Once a stable version of the Domain Model is available, POSTDATA will follow the activities described by Me4MAP. One important activity is the mapping of the Domain Model to a semantic model, where every property of the Domain Model will be matched by an RDF vocabulary term, other constraints such as cardinality, domain and range must also be defined. In order to enhance interoperability, this matching has to be done taing care that the most popular vocabularies of the LOD ecosystem are used. A study of the most used vocabularies will be performed so as to find the best options for each term. Examples of RDF vocabularies that might be used are Dublin Core Metadata Terms, Digitised Manuscripts to Europeana (DM2E), Friend of a Friend (FOAF) vocabularies, among others. This semantic model will be tested in an iterative process of development, using resources from metrical repertoires that were not used as cases in the metadata application profile process development.

Bibliography

Birnbaum, D. J., & Thorsen, E. (2015a). Markup and meter: Using XML tools to teach a computer to think about versification. In *Balisage: The Markup Conference*.

Birnbaum, D. J. & Thorsen, E. (2015b). Enabling the automated identification and analysis of meter and rhyme in Russian verse. In *DH 2015: Global Digital Humanities*, Sydney.

Bosch, Mela & Rio Riande, Gimena del (2016). Las Humanidades Digitales o el ornitorrinco". In *Las Humanidades Digitales en Argentina: discursos y tecnologías en cruce.* Retrieved from: http://www.centrocultural.coop/blogs/utopia/2016/11/01/cronica-de-la-actividad-las-humanidades-digitales-en-argentina/ – accessed 25 March, 2017.

Chatman, S. & Levin, S. R. (1967). *Essays on the language of literature*. Boston: Houghton Mifflin.

Curado Malta, M., & Baptista, A. A. (2013a). A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. In M. Foulonneau & K. Eckert (Eds.), *Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013* (pp. 90–103). Lisbon: Dublin Core Metadata Initiative. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/178/81 – accessed 23 March, 2017.

Curado Malta, M., & Baptista, A. A. (2013b). Me4DCAP V0. 1: A method for the development of Dublin Core Application Profiles. In *Information Services and Use* (Vol. 33, pp. 161–171). Dublin Core Metadata Initiative. Retrieved from http://doi.org/10.3233/ISU-130706 – accessed 23 March, 2017.

Curado Malta, M., & Baptista, A. A. (2013c). State of the Art on Methodologies for the Development of Dublin Core Application Profiles. *International Journal of Metadata, Semantics and Ontologies*, *8*(4), 332–341. Retrieved from http://doi.org/http://dx.doi.org/10.1504/IJMSO.2013.058416 - accessed 23 March, 2017.

Curado Malta, M. (2014). *Contributo metodológico para o desenvolvimento de perfis de aplicação no contexto da Web semântica*. Universidade do Minho, Escola de Engenharia. Programa Doutoral em Tecnologias e Sistemas de Informação. Retrieved from http://hdl.handle.net/10171/35718 – accessed 23 March, 2017.

Curado Malta, M., Centenera, P. & Gonzalez-Blanco, E. (2017). Using Reverse Engineering to define a Domain Model: The case of Development of a Metadata Application Profile for the European Poetry. In Curado Malta, M., Baptista, A. A. and Walk, P. (Eds.), *Developing Metadata Application Profiles* (pp. 146-180). Hershey PA: IGI Global. DOI: 10.4018/978-1-5225-2221-8.ch007.

González-Blanco García E. & Seláf, L. (2014). Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires". In L. Soriano, M. Coderch, H. Rovira, G. Sabaté & X. Espluga (Eds.), *Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world* (pp.321-322). Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien: Peter Lang.

González-Blanco García, E., Rio Riande, G. del, Martínez Cantón, C. & Martos Pérez, M. D. (2014). "La codificación informática del sistema poético medieval castellano, problemas y propuestas en la elaboración de un repertorio métrico digital: ReMetCa". In A. Baraibar (Ed.), *Visibilidad y divulgación de la investigación desde las humanidades digitales. Experiencias y proyectos* (pp.185-203). Pamplona, Servicio de Publicaciones de la Universidad de Navarra. Colección BIADIG (Biblioteca Áurea Digital), 22 / Publicaciones Digitales del GRISO. Retrieved from http://hdl.handle.net/10171/35718 - accessed 23 March , 2017.

González-Blanco, E. G.,Martínez Cantón, C. & Rio Riande, G. del (2015). Making visible the invisible: metrical patterns, contrafacture and compilation in a Medieval Castilian Songbook. DH 2015: Global Digital Humanities, Sydney. Retrieved from http://dh2015.org/abstracts/ - accessed 23 March, 2017.

González-Blanco, E. G., Rio Riande, G. del & Martínez Cantón, C. (2016a). Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires. In John P. McCrae, Christian Chiarcos et al. (Eds.), *Proceedings of the LREC 2016 Workshop, LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-LDL2016_Proceedings.pdf - accessed 25 March, 2017.

González-Blanco, E. G., Rio Riande, G. del & Martínez Cantón, C. (2016b). "DH Poetry Modelling: a Quest for Philological and Technical Standardization". DH 2016, Krakow. Retrieved from http://dh2016.adho.org/abstracts/73 – accessed 23 March, 2017.

Nilsson, M., Baker, T., & Johnston, P. (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved from http://dublincore.org/documents/singapore-framework/- accessed 23 March, 2017.

Pue, S. A.; Teal,T. K. & Brown, C. T. (2015). Using Bioinformatic Algorithms to Analyze the Politics of Form in Modernist Urdu Poetry. DH 2015: Global Digital Humanities, Sydney.

Rumbaugh, J., Jacobson, I. and Booch, G. (2004). T*he Unified Modeling Language Reference Manual* (2nd Edition). Boston: Addison-Wesley.

W3C (2010). RDF - W3C standards. http://www.w3.org/RDF/ - accessed 23 March, 2017.

Vers les Données Liées : Conséquences Théoriques et Pratiques Pour les Sciences Humaines

Lyne Da Sylva
lyne.da.sylva@umontreal.ca
Université de Montréal, Canada

Introduction

Le Web ne sera peut-être jamais plus comme on le connaît maintenant. Ce vaste répertoire de connaissances et informations publiées et commentées par les internautes risque, si la tendance des travaux du W3C se maintient, de s'enrichir d'une couche de représentation supplémentaire. Le Web sémantique et les données liées ne visent pas à remplacer le Web existant, mais à s'y greffer. Ils représentent un nouveau paradigme de représentation de l'information sur le Web, non plus comme des documents cohérents (des pages Web lisibles par l'humain), mais plutôt comme des jeux de données (Linked Open Data) représentant des relations binaires entre un objet et une propriété de celui-ci : des « triplets », encodés selon des standards précis et permettant des traitements automatiques à grande échelle. À la place du Web existant, ou en complément de celui-ci, le Web de données ouvre de multiples possibilités de mise en rapport d'informations diverses. Il fait miroiter des possibilités de recherches d'information plus élaborées (parce que basées sur des ensembles d'inférences éventuellement complexes) et de traitements automatiques sophistiqués sur les données.

Tout ceci est sans doute vrai; plusieurs membres de la communauté scientifique et de diverses communautés professionnelles ont déjà contribué à construire le désormais gigantesque réseau Linked Open Data (Europeana, OCLC, Library of Congress, ACM, BBC Music,...).

L'objectif de cette communication n'est pas de remettre en question le processus, mais bien de poser un regard éclairé sur certaines des conséquences de ce passage des documents aux données liées pour les sciences humaines.

Brève description du Web sémantique et des données liées

Le Web sémantique repose de manière importante sur l 'identification et la description d'entités : les entités dignes d'importance, dans les représentations du Web sémantique, incluent toute entité sur laquelle on peut vouloir exprimer des propriétés ou relations. Ces entités peuvent être des ressources présentes sur le Web (des sites web, des bases de données, etc.) ou non : des personnes en chair et en os, des livres imprimés, des sites géographiques, archéologiques, touristiques, etc. Ou toute autre entité abstraite comme une date, une couleur, un rite funéraire, etc.

Nous présenterons les éléments de base du Web de données : les identifiants (URI) qui représentent les entités, les relations de diverses natures qui permettent de relier les entités, les triplets RDF qui servent à encoder les relations entre entités, les vocabulaires et ontologies qui permettent des descriptions uniformisées, les triplestores qui emmagasinent les jeux de triplets RDF.

Nous soulignerons également les avantages de l'utilisation de données liées en sciences humaines, en illustrant à l'aide de quelques projets de recherche récents (en histoire : Michon, 2016; en littérature : RDA, 2015; en musique : Cannam et al, 2011)

Conséquences pratiques

Suite à cette brève présentation, nous examinerons d'abord les conséquences pratiques du passage aux données liées pour les sciences humaines. Celui-ci influe premièrement sur le focus de la recherche, qui se fait alors graduellement vers l'information et non le document (processus antinomique à certaines disciplines en sciences humaines). Deuxièmement, le travail nécessaire à l'extraction des données est considérable, contrairement à ce qui est de mise en sciences pures (où une partie de l'encodage des données liées est fait à partir de données discrètes obtenues par des appareils de mesure : sondes océanographiques, lectures géologiques, observations astronomiques). Troisièmement, nous verrons que cette transformation rappelle la notion de « redocumentarisation » déjà observée pour le document numérique (Pédauque, 2007). Le terme est utilisé pour décrire les bouleversements induits par l'apparition du document numérique; l'encodage des données dans les technologies du Web sémantique exige d'extraire l'information encodée dans des documents existants et de l'exprimer dans un nouveau format. La création de chaque triplet nécessite un travail d'analyse fine des documents et la décomposition en ses unités élémentaires. On parlera ici de l'atomisation de l'information contenue dans les documents, qui a également des conséquences théoriques (voir ci-dessous). Mais les conséquences pratiques sont déjà considérables.

Conséquences théoriques

Les conséquences théoriques du passage au Web de données touchent d'abord la nature des objets de l'étude. En premier lieu, nous aborderons la réification des éléments d'information : tout énoncé (triplet RDF) destiné au Web de données requiert (i) que l'entité à décrire soit définie ontologiquement dans l'univers de référence (le vocabulaire ou l'ontologie); (ii) que l'entité qui lui est reliée reçoive le même traitement (même lorsqu'il s'agit de concepts abstraits comme la couleur d'un objet); et (iii) que la relation entre les deux entités soit elle aussi réifiée. La réification des deux entités ne choquera pas le chercheur, habitué à scruter l'essence des concepts qu'il étudie. Mais la réification de la relation – l'élévation au rang de concept de tout type de relation comme « est l'auteur de » ou « a correspondu avec » – changera sans doute de manière considérable la vision que le chercheur aura de cette relation.

Une deuxième conséquence théorique du passage au Web de données est le fait que la distinction entre données et métadonnées devient floue, voire inutile. Par le biais de l'extraction de données à partir de documents primaires, tout ce qui est encodé devient métadonnée. Mais si tout est métadonnée, de quelles données sont-elles les métadonnées? Les données disparaissent-elles? ou est-ce plutôt le concept de métadonnées qui disparaît?

Nous discuterons également de l'uniformisation exigée par les vocabulaires (ou référentiels) partagés. Si les vocabulaires sont davantage stables dans les sciences dites exactes, la terminologie fait moins consensus en sciences humaines et sociales (ceci a été largement documenté dans la construction de thésaurus documentaires). On peut s'attendre à un plus grand nombre de référentiels concurrents. Il est possible que le travail sur les ontologies ait un effet important sur la définition des concepts de base des disciplines et les tentatives de rapprochement entre disciplines. Les travaux en terminologie et en conception de thésaurus pourraient être mis à contribution dans l'entreprise. D'un autre côté, le fait que les données soient liées permet des interconnexions (et un potentiel de normalisation) qui n 'est pas requis lorsque les recherches se font davantage en parallèle.

Enfin, nous présenterons la notion de l'atomisation des objets de recherche. L'apport intellectuel des chercheurs en sciences humains est le résultat d'un travail d'analyse, de mise en rapport, d'abstraction à partir de données disséminées (documents historiques, phénomènes linguistiques, observations sur le terrain …). Or ces documents sont des créations, des synthèses, exprimant des idées réunies par l'auteur en propositions, en phrases, en paragraphes cohérents. À partir du moment où ces informations sont atomisées pour le Web de données, ce travail d'analyse est (potentiellement) perdu, disséminé dans les simples triplets. On peut avancer qu'au contraire, les relations identifiées par les chercheurs, et exprimées par les documents résultants, peuvent faire partie des relations exprimées par les triplets RDF. Mais on doit bien prendre conscience du fait que chaque énoncé reste, dans le formalisme, isolé des autres. Reste aux chercheurs futurs la tâche de développer des moyens de faire des abstractions additionnelles (à un niveau plus macro) à partir des simples triplets RDF.

Conclusion

Le passage au Web de données peut être extrêmement utile à la recherche en sciences humaines. Nous présenterons un certain nombre de conséquences pratiques et théoriques de cette mutation. Le passage aux données implique l'atomisation du sujet de recherche et la redocumentarisation de sa matière première, qui exige un travail important d'extraction d'informations et de réencodage. Il a comme conséquence la réification de chaque élément d'information inclus dans la nouvelle description. Les représentations résultantes brouillent la frontière entre métadonnées et données. Enfin, le passage aux données liées entraînera un partage des référentiels, un potentiel de plus grande interopérabilité entre les descriptions et, peut-être, une injection de travaux sur les concepts fondamentaux des disciplines, de pair avec des travaux en terminologie et en vocabulaires contrôlés. Il est important pour les chercheurs en sciences humaines de mesurer l'apport potentiel du Web de données et les transformations qu'il pourra apporter au développement de leur science ; la présente communication veut amorcer la réflexion sur le sujet.

Bibliography

Cannam, C., Sandler, M., Jewell, M.O., Rhodes, C., d'Inverno, M. (2011). Linked Data and You: Bringing Music Research Software into the Semantic Web. *Journal of New Music Research*. Volume 39, no 4: Music Informatics and the OMRAS2 Project.

Linked Open Data. http://linkeddata.org/

Michon, P. (2016). Données liées historiques : De la nécessité d'un partenariat entre l'archivistique et l'histoire. http://congres.archivistes.qc.ca/wp- content/uploads/2016/08/DonneeHistoriques.pdf

Pédauque, R.T. (collectif), (2007), *La Redocumentarisation du Monde*, Paris : Éditions Cépadues.

RDF. (2015). *Jane-athon*. http://rballs.info/topics/p/jane/janeathon.html

Numbers into Notes: digital prototyping as close reading of Ada Lovelace’s ‘Note A’

David De Roure
david.deroure@oerc.ox.ac.uk
University of Oxford, United Kingdom

Pip Willcox
pip.willcox@bodleian.ox.ac.uk
University of Oxford, United Kingdom

Lovelace and Babbage

Ada Lovelace is widely held to be the first computer programmer, composing the first algorithm designed for execution by a general purpose computing machine. The algorithm was published in the “notes by the translator”, appended to her translation of Menebrea’s 1842 French account of Charles Babbage’s 1840 seminar on his proposed steam-powered computer, the Analytical Engine, at the University of Turin (Lovelace, 1842).

This third-remove contribution appearing in small print after the main article belies both the closeness of her friendship with Babbage and how extraordinary her vision was of what such a general purpose computing machine could achieve. The phrase often quoted in the literature of computers and music offers another insight into her imaginative response to the hypothetical Analytical Engine:

“Supposing, for instance, that the fundamental relations of pitched sounds in the science of harmony and of musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent.” (Note A)

Our work described in this paper explores this observation, taking inspiration both from Lovelace’s ideas and from composer Emily Howard’s creative response to them. Using digital technologies, we have prototyped software and hardware to investigate Lovelace’s theory of creative computing. We developed a web-based tool, Numbers into Notes, which we used as a prompt to engage the public with these ideas at live musical performances. The musical outcomes of our research are recorded as Digital Music Objects (De Roure, 2016a) which include provenance records.

Following Galey and Ruecker, we suggest that this application can be viewed as a designed digital artefact analogous to a critical work, explicating our interpretation of Lovelace’s words (Galey, 2010).

Simulator

Babbage’s Analytical Engine remains unconstructed, but today we can simulate the execution of programs on it digitally, based on the detailed accounts of the Engine’s design provided by Babbage and Lovelace. As an experiment in hearing what Lovelace might have imagined, we coded an algorithm to run on a simulator in order to generate number sequences which are then mapped to instruments. The numbers are strictly faithful to nineteenth-century mathematics. Human intervention decides the algorithmic parameters and the mapping of the numbers to notes and instruments, to be explored based on the musical context of the time. This experiment led to the creation of the Numbers into Notes web site, making the tools available to others for investigation and composition.

Performance

Emily Howard has also responded creatively to Note A, and the life of Lovelace, in composing her *Lovelace Trilogy* (Petri-Preis, 2013). ‘Ada sketches’ is a short operatic work for mezzo-soprano, flute, clarinet and percussion. Howard’s time as Composer in Residence at the University of Liverpool’s Department of Mathematics, and her work there with Lasse Rempe-Gillen, led to this piece which has been performed in formats that encourage audience response and participation. A performance at the University of Oxford as part of celebrations to mark Lovelace’s 200th birthday investigated the audience’s reception of the work. Its most recent performance, at the Royal Northern College of Music, used our Numbers into Notes web application.

Digital–physical

The Numbers into Notes software invites a thought experiment: had Lovelace lived longer, and had Babbage successfully built the Analytical Engine, what might have happened in pursuit of Lovelace’s musical observation? We extended this thought experiment to ask “what might Lovelace do today?” To explore this, we constructed multiple physical devices (based on the Arduino open-source electronic prototyping platform) to re-enact the algorithms designed for the Analytical Engine (De Roure, 2016b). Today Lovelace could combine multiple machines, and the computational power would enable real-time synthesis, putting into practice the mathematical notions of consonance that were established in the eighteenth century.

Experimental humanities

This approach, which we suggest might be framed as *experimental humanities*, has attracted engagement during events and online, and has also been a successful vehicle in teaching (in the Social Humanities strand of the Digital Humanities at Oxford Summer School). Our work uses digital tools, co-designing digital and digital-physical artefacts, to explore and re-imagine prospective and theoretical technologies of Lovelace’s day. Through this we provoke new responses and discoveries relating to music practice and performance, and to the philosophy and history of technology. The practice of this *experimental humanities* approach enables critical reflection and re-interpretation: we suggest that the digital artefacts we have produced are each interpretations drawing on the life and writing of Lovelace, and the value of this practice lies in the new insights and works they inspire.

This paper recounts these experiments that play at once into generative design and into alternative histories of algorithms and mechanisms. Through making, through prototyping and co-design, we close-read the thought processes Lovelace and Babbage recorded. We point to paths in the development of computing and programming that were not taken, and extend beyond what was practicable in the nineteenth century. Our work also touches on creativity, as anticipated by Lovelace and recast in the "Lovelace questions" (Boden, 1990), and manifest today in the fields of computational creativity and creative computing.

Acknowledgements

This work is supported by: *Fusing Semantic and Audio Technologies for Intelligent Music Production and Consumption*, funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/L019981/1, a collaboration between Queen Mary University of London, University of Nottingham and University of Oxford; *Transforming Musicology*, funded by the UK Arts and Humanities Research Council under grant AH/L006820/1 in the Digital Transformations programme, in collaboration with Goldsmiths University of London.

Bibliography

Lovelace, A.A. (Trans.) (1842) Sketch of the analytical engine invented by Charles Babbage, with notes by the translator. In Scientific Memoirs, Selected from the Transactions of Foreign Academies of Science and Learned Societies, Vol. 3, 1843, pp. 666-731, volume 3. Richard and John E. Taylor, Red Lion Street, Fleet Street, London. Translation of 'Notions sur la machine analytique de M. Charles Babbage' by Luigi Federico Menabrea, in Bibliothèque universelle de Genève. Nouvelle série 41, pp. 352–76.

De Roure, D., Klyne, G., Page, K. R., Pybus, J., Weigl, D. M., Wilcoxson, M., and Willcox, P. (2016a) Plans and performances: Parallels in the production of science and music. In Proceedings of the 2016 IEEE 12th International Conference on e-Science, IEEE, pp. 185–192: 10.1109/eScience.2016.7870899

De Roure, D. and Willcox, P. (2016b) Numbers in places: creative interventions in musical space & time. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct - Mobile- HCI '16 (2016), Association for Computing Machinery (ACM), pp. 1059–1063: 10.1145/2957265.2964199

Galey, A., Ruecker, S. (2010). How a prototype argues. Literary and Linguistic Computing, (25) 4, pp. 405-24: 10.1093/llc/fqq021

Petri-Preis, A. (2013) Emily Howard's Lovelace Trilogy: a musical homage to a mathematical pioneer. Tempo 67 (265) 28–36: 10.1017/S0040298213000442

Boden. M. A. (1990). The Creative Mind: Myths and Mechanisms. Weidenfield and Nicholson, London.

A Capacity Building Framework for Institutional Digital Humanities Support

Quinn Dombrowski
quinnd@berkeley.edu
UC Berkeley, United States of America

Joan Lippincott
joan@cni.org
Coalition for Networked Information
United States of America

The development of well-known institutional digital humanities programs-- such as those at the University of Virginia, University of Victoria, University of Maryland, and others -- has been significantly influenced by unique combinations of dynamic individuals and receptive institutional circumstances. Faculty and administrators in leadership positions at a growing number of universities are interested in developing some form of institutional support for digital humanities research and/or pedagogical practices, but are starting with a fragmented local landscape. Where previous work has focused on particular facets of successful digital humanities programs (e.g. organizational models in Maron & Pickle 2014), addressed digital humanities within a particular organizational context (e.g. in libraries, Varner & Hswe 2016 and Schaffner & Erway 2014) or provided case studies from a single institution (Maron 2011), to date there has not been a holistic framework that can serve as a guide for improving institutional support.

In spring 2016, the EDUCAUSE Center for Analysis and Research (ECAR) and the Coalition for Networked Information (CNI) convened a working group of fifteen library and IT professionals representing a wide range of institution types in the US and Canada, with the goal of developing a "maturity framework" for institutions that are either getting started on developing a digital humanities support program/center or striving make additional progress with their existing efforts. While the maturity framework model is commonly used in ECAR working group publications to describe the adoption or implementation of new technologies or methodologies, particularly in an IT context, the working group quickly reached a unified decision that "maturity" was the wrong model for the development of a digital humanities program. Instead, the working group adopted a similar structure -- using defined stages of development across multiple parameters-- and termed it a "capacity building framework", in order to acknowledge that different institutions may prioritize different kinds of capacity, and those choices have little bearing on the “maturity” of the program as a whole.

The working group’s recently-published white paper identifies five major facets of developing institutional support for digital humanities, and characterizes for each one three stages in the development of institutional digital humanities support: early, established and high capacity. The paper acknowledges that different institution types, and even individual institutions, will choose different areas to prioritize, and there is no “one-size-fits-all” recommendation.

1. **Governance** -- how, and at what level(s), are decisions made? At the early stage, there is no governance in place beyond the individual project level. As a result, individual projects or courses flourish, or fail to do so, largely based on individual PIs’ personal connections and ability to secure funding and resources. For well-connected and charismatic scholars, the early stage may seem like the ideal one, as there are no systems in place to impede them from making any technical decision that suits them, and involving as many support staff as they can convince to participate. While the early stage model may be conducive to individual project capacity, given a particular kind of PI, it does little to build capacity at the institutional level. The white paper takes the position that governance is an important aspect of institutional capacity-building, by providing structures for transparent decision-making that apply equally to all projects. Depending on the institution, governance structures may be in place to determine allocation of internal grant funding or consulting resources, coordinate the purchase of software licenses or expensive hardware, and/or provide input into larger decisions in program development (e.g. whether or how to offer degree programs, whether to participate in consortial efforts to develop infrastructure or training programs, etc.)

2. **Infrastructure** -- providing access to both technology and expertise -- sits at the core of institutional support programs for digital humanities. The specific tools and resources provided can vary across institutions, depending on the needs and interests of researchers and instructors, as well as the particular skills and expertise of those providing support. The white paper notes that the effectiveness of technical infrastructure is contingent upon the availability of experts who can help scholars make use of that technology. While this is especially true for resources that have a high barrier to entry (such as high performance computing clusters), it also applies to software (e.g. for GIS, 3D modeling, text analysis, and OCR) and more prosaic infrastructure such as web hosting.

3. **Roles and Capabilities** -- successful digital humanities work is most often conducted in teams, with roles and capabilities from three complementary categories: technical experts, champions of engagement, and content innovators. Within these broad categories, there is great deal of overlap and interdependence; and it is through the overlap of roles and interdependence of capabilities that DH can flourish. Depending on the organizational model in place, projects may draw upon skilled collaborators at various places within an institution (e.g. museums and archives, central IT units) or even at other institutions through consortial agreements.

4. **Communication and Outreach** -- in the early stage of institutional support for digital humanities, individual practitioners are not aware of one another, and there are no established channels for disseminating news and announcements about events, workshops, grants, and other opportunities relevant to digital humanities. As institutional support becomes established, digital humanities mailing lists and event calendars are created, funding is accessible for one-off events and activities, and beginner-oriented training becomes available. High capacity for communications and outreach involves coordinated, regular communication, dedicated funding for activities, and multiple levels of training covering a range of skills.

5. **Acceptance** -- i.e. the acceptance of digital humanities work as a component of promotion and tenure, of course assessment, and of performance reviews for librarians and IT staff. The white paper notes that academic acceptance of digital humanities work is significantly influenced by developments within particular disciplines, for example, the development of guidelines such as those produced by the Modern Language Association (2012), American Historical Association (2015), and College Art Association (2016) for evaluating digital humanities scholarship. Nonetheless, an institution can increase its capacity at the local level by having department chairs, deans and provosts publicly take a position supporting the assessment of digital scholarship as part of tenure dossiers and advocate for the consistent application of disciplinary guidelines, where available. Acceptance of digital humanities also applies to the evaluation of courses with digital humanities components, as well as to performance evaluations for librarians and staff who support digital humanities work. In the early stage, librarians and IT staff provide digital humanities

> support “on the margins” of their jobs, whereas at a higher-capacity stage, digital humanities support is an officially recognized aspect of IT staff and librarians’ job descriptions, and is factored into performance evaluations accordingly.

To complement the capacity building framework, the white paper includes a section on getting started with developing institutional capacity. This section has pointers for how to do an environmental scan and needs assessment, a discussion of interdisciplinarity, recommendations for the kinds of partnerships that support institutional capacity-building, and a number of commonly-used organizational models.

Bibliography

American Historical Association (2015). “Guidelines for the Professional Evaluation of Digital Scholarship by Historians”. https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians

College Art Association. (2016). “Guidelines for the Evaluation of Digital Scholarship in Art and Architectural History”. http://www.collegeart.org/pdf/evaluating-digital-scholarship-in-art-and-architectural-history.pdf

Maron, N.L. (2011). “The Department of Digital Humanities (DDH) at King’s College London 2011: Cementing Its Status as an Academic Department”. Ithaka S+R, 2011. http://sr.ithaka.org/?p=22368

Maron, N.L. and Pickle, S. (2014) “Sustaining the Digital Humanities: Host Institution Support Beyond the Start-up Phase”. Ithaka S+R, 2014. DOI: http://dx.doi.org/10.18665/sr.22548

Modern Language Association. (2012) “Guidelines for Evaluating Work in Digital Humanities and Digital Media”. https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-Humanities-and-Digital-Media

Schaffner, J. and Erway, R. (2014) “Does Every Research Library Need a Digital Humanities Center?”. OCLC. http://www.oclc.org/research/publications/library/2014/oclcresearch-digital-humanities-center-2014-overview.html

Varner, S., and Hswe, P. (2016). “Special Report: Digital Humanities in Libraries”. American Libraries Magazine, 2016. https://americanlibrariesmagazine.org/2016/01/04/special-report-digital-humanities-libraries/

Corpora and Complex Networks as Cultural Critique: Investigating Race and Gender Bias in Graphic Narratives

Alexander Dunst
dunset@mail.upb.de
University of Paderborn, Germany

Rita Hartel
rst@uni-paderborn.de
University of Paderborn, Germany

Introduction

This paper reports on efforts to integrate cultural critique into a DH project that analyzes a corpus of book-length comics, or graphic narratives. We argue that the analysis of issues such as gender, race, and class should be central to digital scholarship that aims to become accessible to the public and appear relevant to the humanities at large. Therefore, cultural criticism needs to be integrated into digital projects from the very beginning. Our research takes up calls to “design for difference” and to develop visualizations that “enact [the] humanistic properties” of complexity and contradiction (McPherson and Drucker, 242). Part I looks at the construction of a corpus of graphic novels, memoirs, and non-fiction. Basing our corpus on academic databases, international comics awards, literary histories, and online booksellers provides insight into institutional gender and racial biases, as well as the opportunity to address them. Part II takes up Drucker’s criticism of network analysis as reductive and static. We present networks of a pilot corpus that pay attention to social inequality and replace reductive edges with distinct forms of communication. Part II exemplifies our intention to make DH scholarship relevant to a wider public. The broad appeal of comics presents an ideal point of entry for people who might not otherwise be interested in digital research. We apply the popular Bechdel Test (proposed by Alison Bechdel in a comic strip but used mainly to study film and TV), to highlight the male bias of graphic narratives.

Corpus Analysis of Institutional Gender and Racial Bias

The traditional canon of literary studies has long been criticized for its exclusion of female, non-Western, and minority authors. As a much younger field, comics studies lacks the extensive canons and bibliographies produced by literary historians. This does not mean that similar biases are absent, however. As part of a larger project, we have built a monitor corpus of book-length comics by drawing on sources that include academic databases (JSTOR, MLA), international comics awards, literary and cultural histories of comics, news media coverage, and Amazon.com (Dunst et al.). Of 220 titles included in the corpus by fall 2016, 84 per cent were written by male authors and 73 per cent were identified as white. Biases are unevenly distributed:

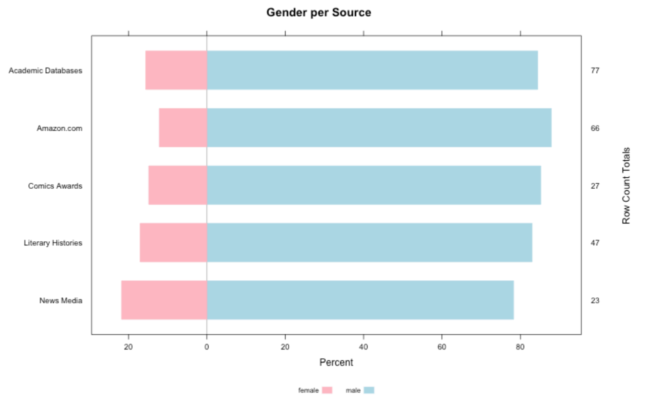


Figure 1: Gender of issue's author grouped by book's source

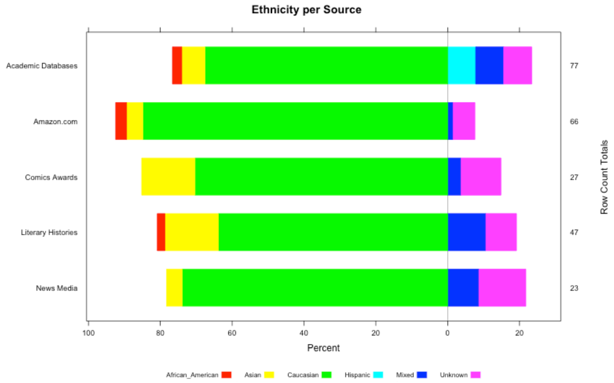


Figure 2: Ethnicity of issue's author grouped by book's source

The absence of reliable bibliographies means that the size, gender and racial make-up of the population remains uncertain. Yet given the considerable differences between sources, institutional biases appear likely. To address these existing biases, two steps were undertaken. A survey sent to comics scholars (five female, five male) asked them to suggest between five and ten graphic narratives written by women that should be included in the corpus. Of a total of 53 suggestions by nine respondents, nine volumes were listed by more than one scholar and 12 had already been included in the corpus, while 14 fell outside of the sampling frame. 16 new works were added, bringing the ratio of female author to slightly less than 22 per cent. The second step includes a comparison of the monitor corpus and collections held at the Library of Congress and the Billy Ireland Cartoon Library at Ohio State University. By checking authors in these collections against a list of names that were assigned genders by the US Social Security Administration, we compare their gender make-ups and will potentially add to our corpus.

Gender and Interaction Types in Semi-Automatic Networks

Network analysis has steadily grown as an area of research since Franco Moretti's visualization of Shakespeare's *Hamlet* (Moretti). Scholars have focused on automatic extraction and statistical analysis of data from novels, plays, and intellectual networks (Elson, Dames & McKeown; van de Camp & van den Bosch). Recent efforts include computing main characters and operationalizing dynamic networks (Jannidis et al; Karsdorp & van den Bosch; Xanthos et al). While these networks answer some of Drucker's criticism, the approaches remain reductive. Limiting interactions to undifferentiated edges appears particularly unsatisfying for visual media, in which communication takes visibly different forms: characters may look at and touch each other, or appear together in a panel. Despite recent advances, computer vision has trouble recognizing non-perspectival drawings and applying OCR to handwritten comics fonts remains fraught with difficulty (Dunst et al; Rigaud 2013 & 2015). As the automatic extraction of network data is some way off for comics, we focus on networks that are semantically enriched via manual annotation to engage with the central questions posed by cultural studies. The following network (Figure 3) of Karasik and Mazzuchelli's *City of Glass* combines different types of interaction with gender assignations:

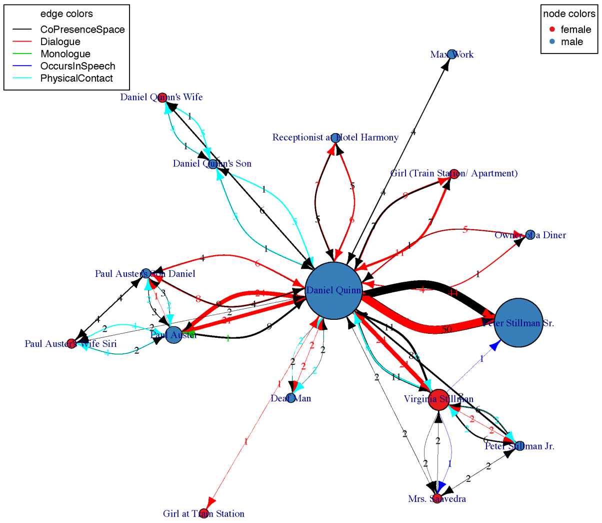


Figure 3: Interactions and gender assignations in *City of Glass*

These networks and the SSA name database allow us to study the relation between authors' gender and its fictional representation. Male characters are consistently more central to works by male authors. Notably, female characters show higher betweenness centrality in narratives written by women, as Figure 4 shows.

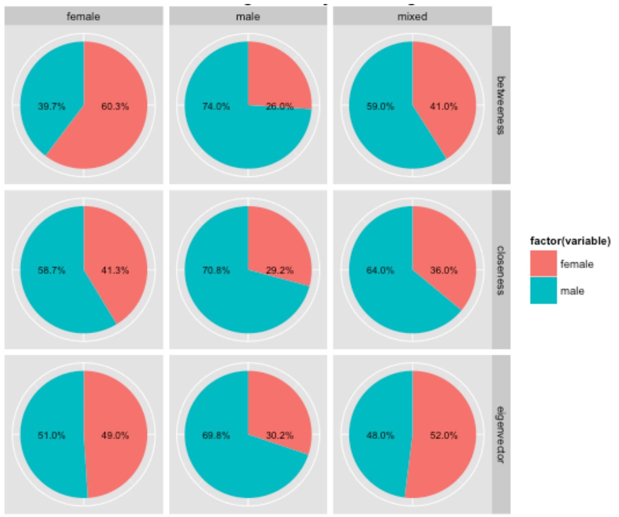


Figure 4: Centralities of character's genders by author's gender

Semi-Automated Bechdel-Test for GNML-Annotated Graphic Narratives

Efforts to automate the Bechdel Test have been limited to plays and film scripts and led to poor results (Lawrence; Agarwal et al). Three conditions need to be met to pass the test: 1. At least two female characters appear in the story. 2. There is at least one conversation between women. 3. The conversation is **not** about a man. Our XML-annotation language GNML, an extension of John Walsh's CBML, allows for automatically checking if graphic narratives fail criteria 1 and 2, and aids evaluation of whether the remaining narratives pass criteria 3. We decided not to rely on error-prone full automation but to use semi-automatic processes that aid human annotators/analyzers in retrieving quality results. GNML annotations contain information on all character occurrences, the gender of a character, and their interactions. As discussed by Agarwal, even sophisticated machine learning approaches lead to unreliable results in deciding whether a conversation centers on a man. Therefore, for criteria 3, we simplify decision-making by providing the annotator with a ranked list of dialogues, based on the number of male names or male personal pronouns per conversation. Significantly, even if a conversation's focus on a male character could be identified automatically, the test would still be error-prone. Conversations may span several panels or pages and automatic separation of these conversations remains difficult

Conclusions and Future Research

We present corpus metadata and semantically-enriched networks of a widely popular but understudied medium that is only beginning to attract attention by DH researchers. These methods are used to analyze gender and racial biases and suggest ways in which DH can appeal to scholars in cultural studies and the wider public. Future work includes expanding pilot studies to cover our entire corpus by integrating advances in OCR and computer vision and thus working towards fully-automatic extraction and analysis. In the meantime, our networks may function as conceptual models exploring how humanistic forms of complexity can be introduced into network analysis. Analyzing and addressing racial biases against minority authors presents hurdles of a different sort. A repeat of our survey for minority authors appears unproblematic but assigning racial identity to names or authors could be viewed as unethical and, given the international nature of our corpus, would have to consider the complex relationship of racial, national, and regional identities.

Bibliography

Agarwal, A., Zheng, J., Kamath, S., Balasubramanian, S., and Ann Dey, S. "Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test". *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*. 830-40

Bechdel, A. (2014) "The Rule". http://dykestowatchoutfor.com/wp-content/uploads/2014/05/The-Rule-cleaned-up.jpg, accessed: 25. 10. 2016

Drucker, J. (2016) "Graphical Approaches to the Digital Humanities". *A New Companion to Digital Humanities*. Ed. S. Schreibman, R. Siemens, & J. Unsworth. Oxford: Wiley. 238-50.

Dunst, A., Hartel, R., Hohenstein, S., and Laubrock, J. (2016) "Corpus Analyses of Multimodal Narrative: The Example of Graphic Narrative". *Conference Abstracts DH 2016*. 178-9.

Elson, D.; Dames, N. and McKeown, K.R. (2010). "Extracting Social Networks from Literary Fiction". *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. 138-47.

Karsdorp, F. and van den Bosch, A. (2016). "The Structure and Evolution of Story Networks". *Royal Society Open Science* 3: 1-15.

Jannidis, F., Reger, I., Krug, M., Weimer, L., Macharowsky, L., Puppe, F. (2016) "Comparison of Methods for the Identification of Main Characters in German Novels". *Conference Abstracts DH 2016*. 578-82.

Lawrence, F. (2011) "SPARQLing Conversation: Automating the Bechdel-Wallace Test". Available at http://nht.ecs.soton.ac.uk/2011/papers/12-flawrence.pdf. (Accessed: 25 October 2016).

Moretti, F. (2011) "Network Theory, Plot Analysis". *Stanford Literary Lab Pamphlets* 2.

McPherson, T. (2014) "Designing for Difference." *Differences* 25:177-188.

Rigaud, C., Karatzas, D., van der Weijer, J., Burie, J.C., and Ogier, J.M. (2013). "Automatic Text Localization in Scanned Comic Books". *International Conference on Computer Vision Theory and Applications 201*3.

Rigaud, C., Guérin, C., Karatzas, D., Burie, J.C., and Ogier, J.M. (2015). "Knowledge-Driven Understanding of Images in Comics Books". *International Journal on Document Analysis and Recognition* 18: 199-221.

Van de Camp, M. & van den Bosch, A. "The Socialist Network". *Decision Support Systems* 53 (2012). 761-69.

Walsh, J. (2012) "Comic Book Markup Language: An Introduction and Rationale". *Digital Humanities Quarterly* 6.

Xanthos, A., Pante, I., Rochat, Y., Grandjean, M. (2016). "Visualizing the Dynamics of Character Networks". *Conference Abstracts DH 2016*. 417-19.

Short samples in authorship attribution: a new approach

Maciej Eder
maciejeder@gmail.com
Institute of Polish Language
Polish Academy of Sciences, Poland

Introduction

The question of minimal sample size is one of the most important issues in stylometry and non-traditional authorship attribution. In the last decade or so, a few studies concerning different aspects of scalability in stylometry have been published (Zhao and Zobel, 2005; Hirst and Feiguina, 2007; Stamatatos, 2008; Koppel et al., 2009; Mikros, 2009; Luyckx and Daelemans, 2011), but the question has not been answered comprehensively. In his recent study, Eder proposed a systematic approach to solve the problem in a series of experiments, claiming that a sample should have at least 5,000 running words to be attributable (Eder, 2015).

The above studies (and many other as well) tacitly assume that there exists a certain amount of linguistic data that allows for reliable authorial recognition, and the real problem at stake is to determine that very value. However, one can assume that the authorial fingerprint is not distributed evenly in a collection of texts. Just the contrary, many experiments seem to suggest that the authorial voice is sometimes overshadowed by other signals, such as genre, gender, chronology, or translation. Some authors, say Chandler, should be easily attributable, while some other authors, say Virginia Woolf, will probably have their fingerprint somewhat hidden. Moreover, authorship attribution is ultimately a matter of context: telling apart Hemingway and Dickens will always be easier than distinguishing the Bronte sisters. On theoretical grounds, then, the minimal sample size can not be determined once and forever for the entire corpus, but may be different for different texts in the corpus.

Method and Data

To scrutinize the above intuition, a controlled experiment has been designed, in which particular text samples were assessed independently (one by one) and compared against the corpus. A following procedure was applied: the entire corpus served as a training set, out of which one text at a time was excluded. This temporarily excluded text was further pre-processed: in many iterations, longer and longer samples of randomly chosen words were excerpted (100 independent samples in each iteration), and then tested against the training set. In each iteration, the total number of correctly “guessed” authorial classes – a single value between 0 and 100 – was recorded, resulting in a row of accuracy scores for a given text as a function of its sample size. The same procedure was repeated for each text in the corpus. The above setup does not need to be supplemented by any cross-validation, because the experiment itself is a variant of a leave-one-out cross-validation scenario. Moreover, each text is re-sampled several times, which can be perceived as an additional way of neutralizing potential model overfitting.

The experiments were repeated a few times. Firstly, three different classification methods have been tested: Support Vector Machines (SVM), Nearest Shrunken Centroids (NSC), and a distance-based learner that is routinely used in authorship attribution tests, namely Burrows’s Delta (Burrows, 2002). However, Delta was used as a general classification framework combined with a few custom kernels that seem to outperform the original setup. These included Cosine Delta (Evert et al., 2016), min-max measure (Kestemont et al., 2016), Eder’s Delta (Eder et al., 2016), and, obviously, the original measure as introduced by Burrows and mathematically justified by Argamon (2011). Secondly, all the tests have been repeated for different vectors of input features, or most frequent words: 100, 200, 300, 500, 750 and 1,000. While the choice of the vectors’ lengths was arbitrary, it was aimed to follow usual stylometric scenarios in their various flavors, ranging from a considerably short list of mostly frequent words, to a longish vectors overwhelmed by content words.

The aforementioned method of testing was applied into two roughly similar corpora (one at a time): a corpus of 100 English novels by 33 authors (male and female), covering the years 1840–1940, and a similar corpus of 100 Polish novels. Both corpora, referred to as the Benchmark Corpus of English and the Benchmark Corpus of Polish, have been compiled by Jan Rybicki (pers. comm.). The corpora used in the experiment, as well as the complete code needed to replicate the study, will be available in a GitHub repository.

Results

A lion’s share of tested samples revealed a very consistent and clear picture. According to intuition, the performance for short samples falls far beyond any acceptance rate, sometimes showing no correct “guesses” at all. This is followed, however, by a very steep increase of performance which immediately turns into a plateau of statistical saturation, despite the number of analyzed features (frequent words). An example of such a behavior is *The Ambassadors* by Henry James (Figure 1), as well as many other novels by Blackmore, Chesterton, Foster, Lytton, Meredith, Morris, Thackeray, and Trollope. As one can see, the amount of text needed for a reliable attribution is less than 2,000 words (!), an amount radically smaller than the previous study suggests (Eder, 2015). Sometimes the picture is somewhat blurry, nevertheless the same general shape reappears, as in the case of *Felix Holt* by

George Elliot (Figure 2). As one can see, using shorter vectors of features requires longer samples to extract the authorial profile.

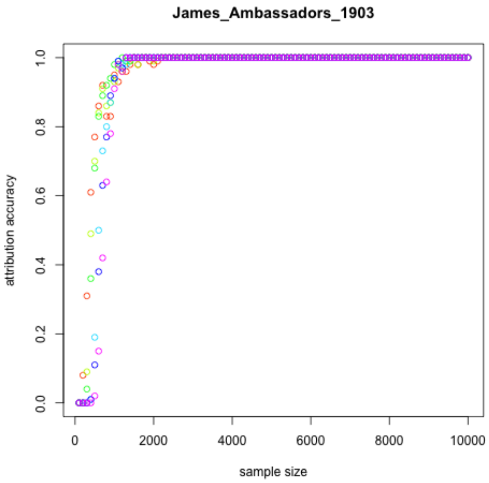


Figure 1: *The Ambassadors* by Henry James contrasted against a corpus of 100 English novels: the attribution accuracy as a function of sample size (in words). Colors represent the results for different vectors of MFWs: 100 (red), 200 (yellow), 300 (green), 500 (cyan), 750 (blue), and 1,000 (violet).

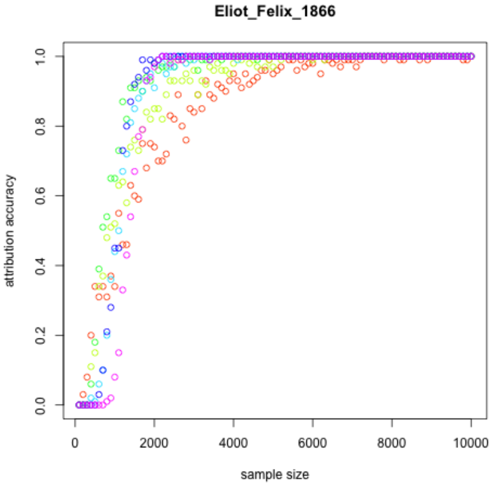


Figure 2: *Felix Holt* by George Eliot: the dependence of authorship recognition and sample size.

Optimistic as they are, however, the results might differ significantly. E.g., in some cases, the statistical saturation does not really take place, even if very long samples are used (Figure 3: scores for *Saints Progress* by John Galsworthy). What is more important, however, the final results additionally depend on the number of analyzed features. In Figure 4, a representative example of this behavior has been shown, namely *Bleak House* by Dickens.

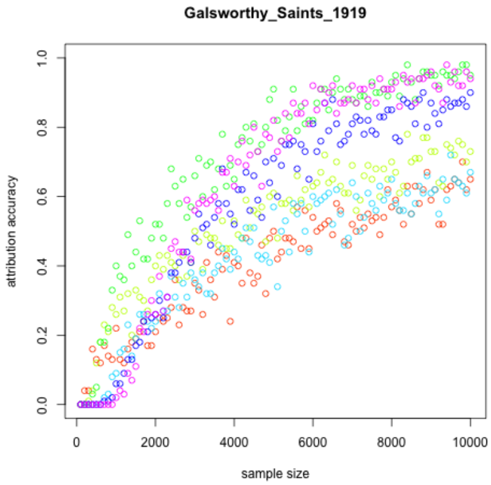


Figure 3: *Saints Progress* by John Galsworthy: the dependence of authorship recognition and sample size.

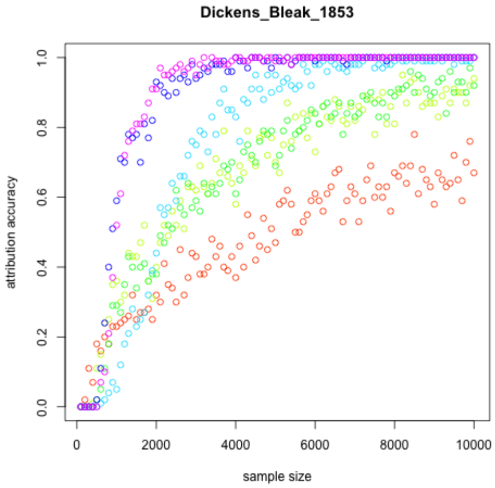


Figure 4: *Bleak House* by Charles Dickens: the dependence of authorship recognition and sample size.

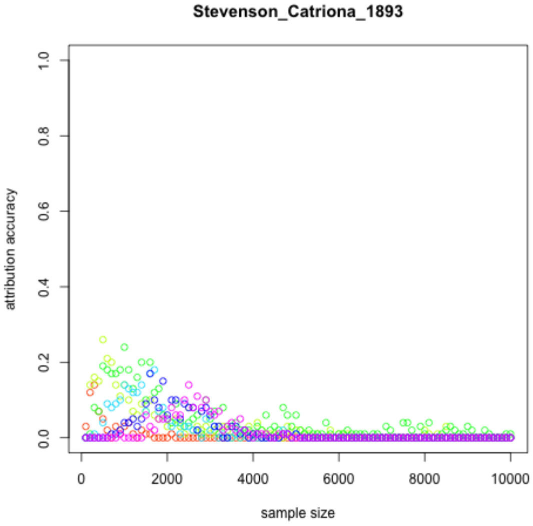


Figure 5: *Catriona* by Robert Louis Stevenson: the dependence of authorship recognition and sample size.

Last but definitely not least, there are a few texts that are never correctly attributed, no matter how long the extracted samples are (Figure 5). The question why some novels were misclassified will be addressed in a separate study. Here, it should be emphasized that such a behavior is unpredictable. Certainly, it can be easily detected, as long as one tests novels of known authorship; it becomes an obstacle, however, when one tries to scrutinize an anonymous text.

Detecting Outliers

The outcome of the above experiment shows that the minimal sample size can be lowered substantially, from ca. 5,000 running words as suggested previously (Eder, 2015), to less than 2,000 words. However, this is true only for those texts that exhibit a clear authorial signal; otherwise the risk of severe misclassification appears. To take advantage of the above results, then, one has to be sure which category an analyzed text belongs to. In a controlled experiment, the task is simple, in a real-case attribution study, however, one has no chance to fine-tune the model by testing the disputed sample against the corpus. What if an anonymous text does not reveal a clear accuracy curve, as the one in Figure 1?

To overcome the sample size issue of unknown texts, an additional measure can be involved to supplement the accuracy scores. (Due to limited space in this abstract, a compact outline of the proposed solution will be presented, rather than a complete algorithm). In the case of misclassification, one would like to know if the wrong response is consistent, or if different classes were assigned chaotically. To address this question, an indicator of consistency would be useful. The Simpson index is a very simple measure of concentration when observations are classified into a certain number of types (Simpson, 1949):

λ = Σpi2

where pi is the proportion of observations belonging to the ith type. The index can be easily adopted to indicate imbalance between assigned classes in supervised classification. To this end, the obtained classification scores (for a given sample size) have to be divided by the total number of trials (in this case, 100). The value 1 reflects purely consistent results, lower values mean that the assigned classes were fuzzy.

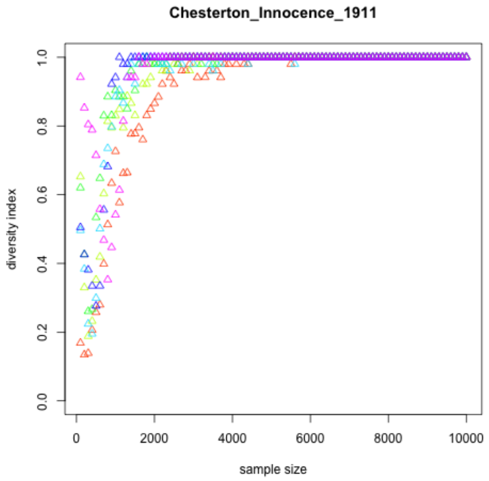


Figure 6: Diversity scores (Simpson index) as a function of sample size.

To make a long story short: the texts that distribute their accuracy curves as in Figure 1 will also exhibit the same shape of the diversity index (see Figure 6). However, when the accuracy scores are low and/or ambiguous, the diversity index might provide a priceless hint. It is especially important when the accuracy scores are consistent (Figure 5), and the Simpson index is not (Figure 7). Instead of being mislead ("Stevenson did not write Catriona", which is not true), we are warned that the classification is inconsistent. Thus, to reliably test a minimal size of a disputed text, one has to take into account two values (accuracy and diversity). The bigger the dispersion between the indices, the smaller the probability that the text is attributable – perhaps a longer sample has to be involved, or a different set of features?

Conclusion

The study was aimed at re-considering the minimum sample size for reliable authorship attribution. The results of the experiments suggest that a sufficient amount of textual data may be as little as 2,000 words in many cases. However, sometimes the authorial fingerprint is so vague, that one needs to use substantially longer samples to make the attribution feasible. A question of some importance is to which category an unknown (disputed) text belongs.

Bibliography

Argamon, S. (2011). Interpreting Burrows's delta: geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2): 131–47.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities,* 30(2): 167–82.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal,* 8(1): 107–21.

Evert, S., Jannidis, F., Proisl, T., Thorsten, V., Schöch, C., Pielström, S. and Reger, I. (2016). Outliers or key profiles? Understanding distance measures for authorship attribution. *Digital Humanities 2016: Conference Abstracts.* Kraków: Jagiellonian University & Pedagogical University, pp. 188–91.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4): 405–17.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems With Applications*, 63: 86–96.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9–26.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1): 35–55.

Mikros, G. K. (2009). Content words in authorship attribution: an evaluation of stylometric features in a literary corpus. In Köhler, R. (ed), *Studies in Quantitative Linguistics,* vol. 5. Lüdenscheid: RAM, pp. 61–75.

Rybicki, J. and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words?. *Literary and Linguistic Computing,* 26(3): 315–21.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163: 688.

Stamatatos, E. (2008). Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2): 790–99.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Proceedings of the Second Asia Conference on Asia Information Retrieval Technology.* (AIRS'05). Berlin, Heidelberg: Springer-Verlag, pp. 174–89.

De la chaîne éditoriale à la plateforme de recherche : structurer, enrichir et diffuser de vastes collections numérisées de publications scientifiques

Nathalie Fargier
nathalie.fargier@persee.fr
ENS de Lyon, France

Brief Summary

Persée développe une plateforme de numérisation, d'enrichissement, de diffusion et d'archivage de publications francophones en SHS, de la 1ère parution à la période la plus récente (revues scientifiques, actes, livres). Résultat d'un travail collégial entre chercheurs, éditeurs, bibliothécaires et ingénieurs, cette plateforme diffuse près de 650 000 documents scientifiques en open access. Elle constitue une expérience originale et concrète pour explorer les enjeux croisés de la numérisation patrimoniale, de l'édition électronique et de l'analyse sémantique. La communication vise à décrire le passage de documents imprimés (représentation arborescente) à une publication numérique puis à un ensemble de données structurées (représentation réticulaire sous forme de graphe) et, à analyser la recontextualisation opérée et les transformations à l'œuvre. La bibliothèque numérique, l'entrepôt OAI et le triplestore Persée sont des points d'accès complémentaires et offrent des potentialités radicalement nouvelles de fouille de texte et de données, d'analyse, de mise en relation et de réutilisation.

Full abstract

Persée est une plateforme de production, d'enrichissement, de diffusion et d'archivage de collections de publications francophones en sciences humaines et sociales, de la 1ère parution à la période la plus récente (revues scientifiques en 1er lieu mais aussi actes de colloques et livres). Résultat d'un travail collégial entre chercheurs, éditeurs, bibliothécaires et ingénieurs, cette plateforme diffuse actuellement plus de 650 000 documents scientifiques en texte intégral et en *open access* et, elle poursuit son enrichissement.

Une décennie après l'ouverture de persee.fr, alors que la numérisation du patrimoine documentaire est massive et que de nouveaux modes de publication ont émergé, cette plateforme constitue une expérience originale et concrète permettant d'explorer les enjeux des collections numériques dans un monde connecté et les interactions entre numérisation patrimoniale et édition électronique. Dans le cadre de cette présentation, il est proposé de décrire les conditions du passage du monde de l'imprimé à celui du numérique, de volumes papier à un ensemble de documents structurés et reliés entre eux et, d'analyser la « *recontextualisation* » qui a été ainsi opérée et les transformations structurelles qui sont à l'oeuvre.

Méthode de fragmentation et chaîne opérationnelle de traitement

Confrontés à des objets qui disposent d'une matérialité évidente (la revue comme un ensemble de documents papier) et d'une pertinence intellectuelle (la revue en tant qu'objet éditorial), nous avons assuré une modélisation des données pour mettre en évidence les structures des documents et leurs dépendances. Ces structures sont représentées dans un modèle qui contrôle la validité des documents. La chaîne éditoriale Persée réunit un ensemble de méthodes et d'outils permettant de constituer des corpus numériques en XML, de manière intégrée et largement automatisée, en vue de leur indexation fine, leur diffusion et leur archivage. Les métadonnées sont encodées selon les schémas Dublin Core, MarcXML et MODS. Le texte intégral issu de la ROC (**reconnaissance optique de caractères**) est disponible selon le schéma TEI et enfin, l'ensemble des données est décrit et organisé au sein d'un container XML au format METS. La publication des documents s'opère par des algorithmes de transformation qui s'appuient sur le modèle pour publier des documents dans des formats standards.

De l’arborescence au réseau : l’évolution de la représentation logique des documents

Les publications scientifiques imprimées sont organisées selon une logique arborescente avec, pour les revues, par exemple, un agencement de type : titre/année/tome/volume/numéro/article. Cette représentation abstraite fait l’objet d’une retranscription numérique jusqu’à un niveau plus précis qui est celui du plan des articles, de la liste des illustrations, des résumés et, des annexes, etc. Considéré isolément et intrinsèquement, l’article devient le nœud d’un nouveau maillage en se fondant sur l’exploitation des citations, des noms d’auteurs et des informations en son sein même. Nous mettons en œuvre un référencement croisé (cite / cited by) et un alignement sur des référentiels d’autorités Auteurs et des sources comme DBpedia. Une analyse plus fine des articles permet d’identifier des entités nommées et d’établir des liens avec des thesaurus disciplinaires. Loin de tout traitement massif, la méthodologie retenue combine des algorithmes de recherche ciblée soumis à validation humaine afin de garantir pertinence et qualité. Les objets numériques ainsi créés se distinguent fondamentalement du matériau papier de base et ils offrent des potentialités majeures en termes de recherche, de Text and Data Mining, d’analyse et de mise en relation. Ainsi, le numérique permet-il de référencer parallèlement la collection appréhendée comme un objet intellectuel à part entière, des segments constitutifs (article/communication/chapitre) et des données, de multiplier les points d’accès à un ensemble structuré, contextualisé et historicisé.

Ouverture de l’accès, ouverture du code source et ouverture des données

Dès l’origine, la voie de l’*Open Access* a été retenue pour la diffusion sans aucune restriction des métadonnées et des documents en texte intégral, l’ouverture et le partage étant considérés comme des instruments essentiels de visibilité et de circulation de la production scientifique dans un environnement où la langue française n’est plus en situation d’hégémonie. Selon une suite logique, les développements informatiques ont été opérés dans un esprit *Open Source* et les données Persée sont intégrées au web de données sous la forme de triplet RDF. L’objectif poursuivi est double : favoriser la réutilisation des contenus dans d’autres contextes que ceux qui ont vu leur création et l’accès de tous au patrimoine scientifique au-delà des cercles académiques.

Bibliography

Babeu, A. (2011). « Rome Wasn't Digitized in a Day »: *Building a Cyberinfrastructure for Digital Classicists*. London : CLIR Publication. 307p

Bachimont, B. (2007). L’indexation multimédia : description et recherche automatique. Paris : Hermès. *Nouvelles tendances applicatives : de l’indexation à l’éditorialisation.* P15-29

Pédauque, R. T. (2006) *Le document à la lumière du numérique. Forme, texte, medium : comprendre le rôle du document numérique dans l’émergence d’une nouvelle modernité.* Caen : C&F Editions. 218p

Salaün J M. (2007) « La redocumentarisation, un défi pour les sciences de l'information ». *Études de communication,* Num. 30, p13-23.

Vitali Rosati, M. (2016). What is editorialization? Sens public. Mars 2016 http://www.sens-public.org/article1059.html

The Seven Words of the Virgin: Identifying change in the discourse context of the concept of virginity in Early Modern English

Susan Fitzmaurice
s.fitzmaurice@sheffield.ac.uk
University of Sheffield

Marc Alexander
marc.alexander@glasgow.ac.uk
University of Glasgow

Justyna Robinson
justyna.robinson@sussex.ac.uk
University of Sussex

Michael Pidd
m.pidd@sheffield.ac.uk
University of Sheffield

Iona Hine
i.hine@sheffield.ac.uk
University of Sheffield

Seth Mehl
seth.mehl.10@ucl.ac.uk
University of Sheffield

Fraser Dallachy
fraser.dallachy@glasgow.ac.uk.
University of Glasgow

Matthew Groves
m.i.groves@sheffield.ac.uk
University of Sheffield

Kathryn Rogers
k.m.rogers@sheffield.ac.uk
University of Sheffield

Brian Aitken
brianaitken@glasgow.ac.uk
University of Glasgow

Introduction

The Linguistic DNA project (LDNA) is an AHRC-funded collaborative project (AHRC grant AH/M00614X/1) between the universities of Sheffield, Glasgow, and Sussex which is designing automatic processes to investigate the emergence and development of concepts in pre-1800 CE print. Employing Early English Books Online, manually-transcribed through the Text Creation Partnership (EEBO-TCP) as its primary dataset, supplemented by Eighteenth Century Collections Online (ECCO-TCP) and other high-quality 18th-century text collections, the project is developing and refining a processing pipeline which assembles groupings of words bound together by their contextual use in printed discourse. The project is charting development of these discourse-embedded word groups across time, investigating how they are shaped by historical and literary contexts, the boundaries and overlap between the groupings, and the interaction of 'encyclopedic' groupings with more traditional 'thesaurus'-style semantic fields.

This paper discusses results from a branch of the project which is investigating incidences of rapid change in the size of semantic categories as represented in *The Historical Thesaurus of English* (Kay et al., 2016). Development of concepts through size of *Thesaurus* categories has been investigated previously (cf. Alexander and Struan, 2013; Jürgen-Diller, 2014), although the extra dimension provided by the outputs of the LDNA processor allows a dramatic leap forward in such research by enabling identification of instances in which change in discourse-embedded word groupings acts as catalyst for corresponding rapid change in the semantic fields of English.

The present case study investigates words relating to the concept of 'virginity', utilising processed time-slice subsets of EEBO-TCP as snapshots of the discourse context for these words in Early Modern English print. By building sample word-groupings (the term 'cluster' is here avoided to avoid confusion with cluster or network analysis word clusters) for each of the subsets, it establishes the discourse context of 'virginity' words at different points in the timespan covered by EEBO-TCP. Comparison of these groupings suggests change in focus of language users, through which a largely religious context of use opens out to a secular and then a poetic literary context, suggesting that society's consciousness of this concept and the scale on which it was discussed enlarged dramatically in the period covered by the sub-corpora.

Methodology

In order to select a *Historical Thesaurus* category for analysis, an average pattern of change over time was established. The *Thesaurus* arranges the words of the English language into a semantic hierarchy that is seven category levels deep with the potential for up to four further sub-category levels within any given category. Owing to the incredibly fine-grained nature of the sense categorisation in the *Thesaurus*, it was necessary to 'cut' the hierarchy at human scale using a thematic category set, developed during the AHRC- and ESRC-funded SAMUELS project (Grant AH/L010062/1), which is intended to allow *Thesaurus* users to find information at a level that is salient to human beings – i.e. neither too general nor too detailed.

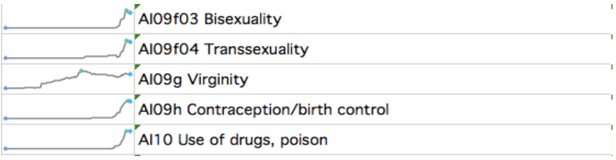

AI09f03 Bisexuality
AI09f04 Transsexuality
AI09g Virginity
AI09h Contraception/birth control
AI10 Use of drugs, poison

Figure 1: Sparkline showing growth of 'Virginity' category from 1000 CE to 2000 CE in context of surrounding thematic categories (which are themselves unusual as they are lexicalised only in later periods of English)

The number of lexemes within each category level was counted, and lexemes were filtered to include only those active within the approximate time range of the EEBO-TCP collection, i.e. 1475-1700 CE. This data was aggregated so that the change in the mean contents of a category could be viewed across time, and decade-to-decade percentage changes calculated. Individual categories were then compared to this average category change, and a deviation of more than 5% from the average change considered to be significant. Out of the categories which were marked as statistically unusual from this process, category 'AI09g Virginity' was selected as promising because the items in its lexis had a relatively low number of homographs that could skew the results towards irrelevant information.

Testing of the LDNA processor outputs is being conducted on select subsets extracted to provide snapshots across the EEBO time-period. The subsets used for this paper cover the periods 1520-39, 1550-59, 1610-11, and 1649. They are designed to contain a similar number of tokens; the progressively contracting timespans reflect the concomitant growth of printed material throughout the 15th to 17th centuries. Each token in the text is regularised, lemmatised, and tagged with a NUPOS part of speech tag via the MorphAdorner pipeline developed by Martin Müller and Philip Burns (Burns, 2013). Data is then gathered by the LDNA processor for the token's co-occurrences within 100- and 200-word bi-directional windows which are intended to simulate paragraph-like sections of the proximate discourse (cf. Fitzmaurice *et al.* forthcoming). Pointwise Mutual Information (PMI) is used to provide a statistic for likelihood of word co-occurrences; a minimum PMI value of 0.5 was arrived at experimentally for identifying node-collocate pairs to be considered interesting in initial stages of investigation.

Seven items – 'maid,' 'maiden,' 'maidenhead,' 'undefiled,' 'vestal,' 'virgin,' and 'virginity' – in the 'Virginity' category were found to be present consistently across the subsets (although an eighth – 'virginal' – was present in the 1520-

39 and 1610-11 subsets). The co-occurrences were then processed to identify those which occurred with multiple items in this list. Words which co-occurred with four or more items were investigated further.

Results

Comparison of the co-occurrence results across the five text subsets shows a consistent shift in the patterns of word association with 'Virginity' category items. The words 'woman' and 'widow' remain strongly associated with the terms across all the subsets, demonstrating societal preoccupation with female rather than male virginity. The most evident change in the grouping is movement from a predominately religious discourse context into the secular world. In the 1520-39 subset, the Virgin Mary is intimately related to discussion of virginity. In the shared collocates listing, *mother* collocates with all seven of the 'Virginity' lexemes, 'mary' with six, 'angel,' 'bless,' 'hymn,' 'nativity,' and 'nazareth' with five each. Of these, only 'mother' maintains a strong association with 'Virginity' words throughout the EEBO period, appearing with four items in the 1610-11 text set and five in 1649.

The secularisation of the term is suggested by the prevalence in later subsets of words relating to marriage, reflecting what appears to be a growing focus on wedlock being preceded by virginity. 'Marry' gradually increases its association with the node items, collocating with four, then five, then seven from 1520 to 1649. 'Marriage' and 'wife' both enter the shared collocate group in 1550, and remain there through to 1649, whilst 'matrimony' is present in 1550, drops out in 1610, and returns in 1649.

The extensive list of shared collocates in the 1649 subcorpus strongly reflects the greater prevalence of literary fiction and poetry in printers' output and reinforces that virginity is a topic for which the discourse context is expanding; where it was easy to intuitively group 'marry,' 'marriage,' and 'wife' together, the 1649 collocates do not form easily identifiable groupings.

Discussion

The consistency of the core items found in the subsets is interesting in its disparity with the *Thesaurus* data, where the increase in the number of terms present in the 'Virginity' category suggests that there should be an expanding number of items found throughout these subsets. The most likely explanation for this is loss of low frequency information through a combination of cut-off values intended to reduce noise for later clustering experiments, and difficulty in normalising/lemmatising low frequency items. A clear outcome of the analysis is the confirmation that the category of 'Virginity' contains core vocabulary which remains almost unchanged in over a century (i.e. 1520-1649), primarily a consistent group of seven items which co-occur with 'Virginity' category words.

This study demonstrates that understanding of semantic development can be enriched by such cross-analysis of discursive-concept word groups with *Thesaurus* semantic fields and the word groups which travel through time with multiple items of *Thesaurus* categories.

Bibliography

Alexander, M. and Struan, A. (2013). "'In Countries so Unciviliz'd as Those?': The language of incivility and the British experience of the world." In Farr, M. and Guégan, X. (eds), *Experiencing Imperialism: The British abroad since the Eighteenth Century, vol. 2*. London: Palgrave Macmillan. pp. 232-249.

Burns, P. R. (2013). *MorphAdorner v2: A Java library for the morphological adornment of English language texts*. Evanston, IL. Northwestern University. http://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf. (last accessed on 31st March, 2017).

Diller, H-J. (2014). *Words for Feelings*. Anglistische Forschungen vol. 446. Heidelberg: Universitätsverlag Winter

Fitzmaurice, S., Robinson, J., Alexander, M., Hine, I., Mehl, S. and Dallachy, F. (forthcoming). "Linguistic DNA: Investigating conceptual change in early Modern English discourse." *Studia Neophilologica*.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I. and Alexander, M. (eds) (2016). *The Historical Thesaurus of English*, version 4.2. Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk/.

De quoi est-il question dans le discours en art contemporain? La fouille de textes appliquée à l'art contemporain dans les centres d'artistes

Dominic Forest
dominic.forest@umontreal.ca
Université de Montréal, Canada

Vinh Truong
congvinhtruong@gmail.com
Université de Montréal, Canada

Yvon Lemay
yvon.lemay@umontreal.ca
Université de Montréal, Canada

Introduction

Cet article présente les résultats d'une recherche dans le cadre de laquelle nous nous sommes intéressés à l'évolution thématique de l'art contemporain dans les centres d'artistes autogérés du Québec. Nous dressons un portrait du milieu de l'art et de son discours en analysant automatiquement un ensemble de textes d'expositions. Notre approche repose sur des techniques automatisées de fouilles de textes. Notre objectif est donc de cartographier l'évolution des thématiques abordées dans le discours en art contemporain.

Notre approche s'inspire de la notion de *distant reading* développée par Franco Moretti. Selon Moretti, l'analyse des textes devrait être effectuée non pas sur une sélection d'ouvrages (le "canon"), mais sur l'ensemble le plus exhaustif possible de documents. Un échantillon de documents trop limité ou trop spécifique risque d'engendrer une perte d'informations précieuses.

Dans le domaine des arts, K. Bender s'est inspiré des travaux de Moretti afin de développer la théorie du *distant viewing.* Appliqué en histoire de l'art où les monographies se concentrent souvent sur les grands artistes et aux grandes oeuvres, il importe de développer, selon Bender, un "regard" permettant de saisir et d'analyser un très grand ensemble d'oeuvres d'artistes. Bien qu'il existe des techniques informatiques permettant d'analyser des images et en extraire des mots-clés descriptifs, ces techniques ne font qu'une description visuelle de l'image.

L'image ou l'oeuvre d'art, qu'elle soit figurative ou abstraite, ne se décode pas simplement. L'art visuel nécessite une analyse qui ne se limite pas au niveau descriptif. Interpréter une oeuvre d'art nécessite à priori certaines connaissances, notamment sur le travail et la démarche de l'artiste, le contexte socio-historique de l'oeuvre ou sur l'histoire et la théorie de l'art. Dans ce contexte, le texte d'exposition constitue un médium riche à exploiter pour extraire de l'information sur les oeuvres.

Le texte d'exposition

Les textes d'exposition en art ont comme objectif de situer et de contextualiser les oeuvres. Ils permettent une meilleure compréhension des oeuvres, du processus créatif de l'artiste, des thèmes, ainsi que des questionnements et des idées évoqués dans les expositions. Dans ce contexte, l'utilisation de techniques de fouille de textes, et plus spécifiquement de modélisation thématique (*Topic Modelling*) (Blei et Lafferty, 2009), permet d'extraire automatiquement les thématiques à partir de textes d'exposition qui n'auraient pu être extraites automatiquement à partir du contenu visuel et qui, à grande échelle, seraient beaucoup difficiles à identifier manuellement. En analysant tous les textes d'expositions disponibles pour une période donnée, nous pouvons poser un regard différent sur le discours en art. L'application de la modélisation thématique sur des textes d'exposition permet, en effet, d'extraire et de structurer automatiquement les principaux thèmes d'un corpus et d'en décrire le vocabulaire spécifique, en plus de permettre la classification d'oeuvres thématiquement comparables. (indépendamment de leur qualités visuelles).

Modélisation thématique et *Artspeak*

Le langage propre au milieu de l'art contemporain, parfois appelé le *Artspeak* ou *International Art English* (Levine et Rule, 2013), peut être difficile à décoder pour les non-initiés. Il a cependant comme qualité de permettre l'expression d'idées et de concepts abstraits propres à l'art contemporain, afin de lexicaliser la complexité des oeuvres. Comme le mentionne Levine, "IAE is about trying to create a more sensitive language, acknowledging the realities of how things [made by artists] work."

La modélisation thématique la complexité des textes ou au niveau d'abstraction des phrases. Il permet, par simple calcul de distribution des mots, de regrouper des textes portant sur des thématiques communes. La modélisation thématique est fréquemment réalisée par l'algorithme LDA (*Latent Direchlet Allocation*), développé David M. Blei, Andrew Ng et Michael I. Jordan en 2003. Bien qu'il ait été développé, il y a plusieurs années, cette algorithme fait l'objet d'un nombre croissant d'utilisation dans le traitement des textes en humanités numériques (Graham, Milligan et Weingart, 2016). Le principe qui sous-tend cet algorithme est qu'un corpus de textes est composé de nombreux thèmes ("topics") latents; ces thèmes étant exprimés par les différents mots qui le composent. Les thèmes sont donc des regroupements de mots effectués sur la base d'un calcul de probabilité (bayésienne). Au niveau individuel, un document est composé d'un nombre *k* de thèmes, lesquelssont composés d'un nombre *x* de mots représentés selon différentes proportions. Au niveau macro (au niveau du corpus en entier), tous les documents d'une collection partagent les mêmes thèmes. L'algorithme LDA présuppose donc l'existence des thèmes latents qui n'apparaissent pas explicitement dans les documents. Ce principe évoque ainsi l'idée d'une structure cachée du document qui nous permet de dégager une structure de mots sur la base de leur proximité thématique. Les mots du corpus sont tous regroupés dans différents thèmes, tout en ayant la possibilité de se retrouver dans plus d'un thème à la fois.

Expérimentation

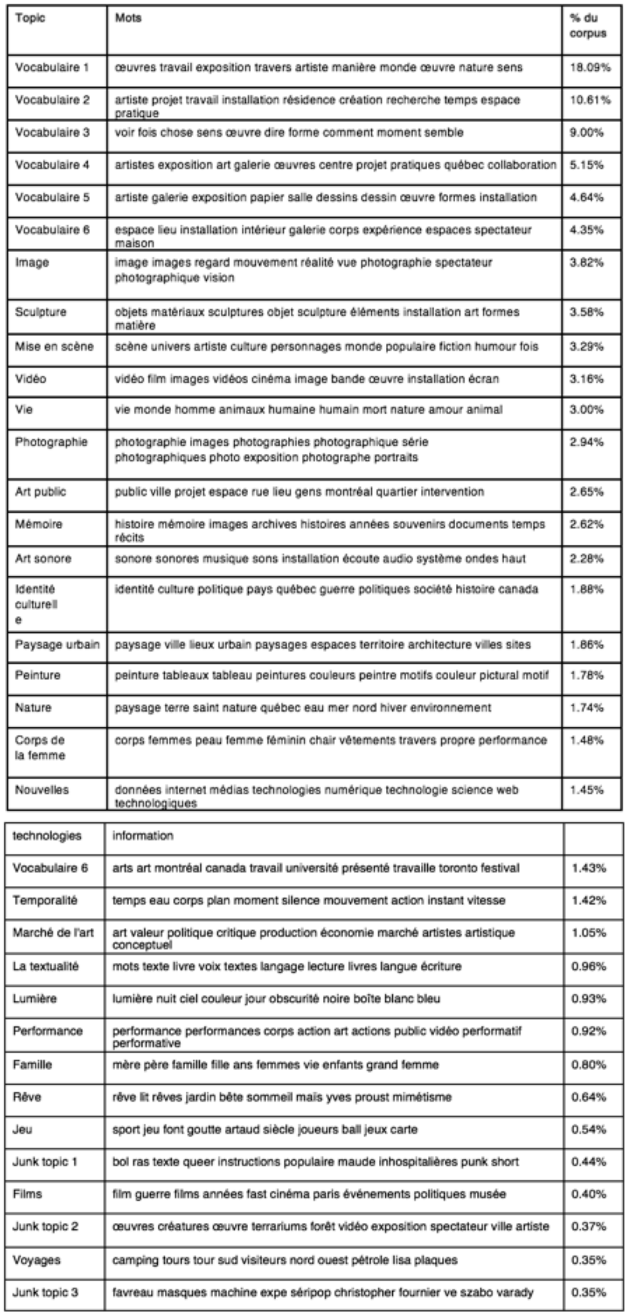
À partir d'une liste exhaustive de galeries d'art membres du RCAAQ (Regroupement des centres d'artistes autogérés du Québec), nous avons constitué un corpus composé de tous les textes d'expositions disponibles sur les sites Web des centres d'artistes (nous définissons ici "exposition" de manière peu restrictive. Il peut donc s'agir de performances, d'événements hors site ou d'expositions individuelles ou collectives). Un total de 3 867 textes d'expositions ont ainsi été trouvés en ligne, datant de 1973 à 2016 et provenant de 49 centres d'artistes autogérés différents. Les données ont été soumis à un processus de modélisation thématique dans l'application *Mallet*.

Divers prétraitements ont été appliqués au corpus initial afin de n'en retenir que le vocabulaire spécialisé en art (suppression des mots fonctionnels, filtrage par seuils statistiques, etc.). Par la suite, le corpus a été soumis à l'algorithme LDA de manière itérative, tout en modifiant le nombre de thèmes afin d'observer les différences dans les thèmes extraits. Le défi intrinsèque à cette approche réside dans le choix du nombre optimal de thèmes à extraire à extraire. Ce paramètre doit être déterminé manuellement.

Nous avons donc analysé un par un les thèmes automatiquement extraits et jugé de leur cohérence, en retournant et en remettant en contexte les termes extraits à l'intérieur des textes originaux. Nous avons finalement opté pour un total de 35 thèmes, nous permettant ainsi d'obtenir un maximum de thèmes cohérents et un minimum de thèmes peu significatifs. Les thèmes ont été étiquetés manuellement à partir des mots les plus représentatifs de chaque regroupement.

Résultats

Le tableau 1 et la figure 1 présente les résultats générés par l'algorithme de modélisation thématique.



Topic	Mots	% du corpus
Vocabulaire 1	œuvres travail exposition travers artiste manière monde œuvre nature sens	18.09%
Vocabulaire 2	artiste projet travail installation résidence création recherche temps espace pratique	10.61%
Vocabulaire 3	voir fois chose sens œuvre dire forme comment moment semble	9.00%
Vocabulaire 4	artistes exposition art galerie œuvres centre projet pratiques québec collaboration	5.15%
Vocabulaire 5	artiste galerie exposition papier salle dessins dessin œuvre formes installation	4.64%
Vocabulaire 6	espace lieu installation intérieur galerie corps expérience espaces spectateur maison	4.35%
Image	image images regard mouvement réalité vue photographie spectateur photographique vision	3.82%
Sculpture	objets matériaux sculptures objet sculpture éléments installation art formes matière	3.58%
Mise en scène	scène univers artiste culture personnages monde populaire fiction humour fois	3.29%
Vidéo	vidéo film images vidéos cinéma image bande œuvre installation écran	3.16%
Vie	vie monde homme animaux humaine humain mort nature amour animal	3.00%
Photographie	photographie images photographies photographique série photographiques photo exposition photographe portraits	2.94%
Art public	public ville projet espace rue lieu gens montréal quartier intervention	2.65%
Mémoire	histoire mémoire images archives histoires années souvenirs documents temps récits	2.62%
Art sonore	sonore sonores musique sons installation écoute audio système ondes haut	2.28%
Identité culturell e	identité culture politique pays québec guerre politiques société histoire canada	1.88%
Paysage urbain	paysage ville lieux urbain paysages espaces territoire architecture villes sites	1.86%
Peinture	peinture tableaux tableau peintures couleurs peintre motifs couleur pictural motif	1.78%
Nature	paysage terre saint nature québec eau mer nord hiver environnement	1.74%
Corps de la femme	corps femmes peau femme féminin chair vêtements travers propre performance	1.48%
Nouvelles technologies	données internet médias technologies numérique technologie science web technologiques information	1.45%
Vocabulaire 6	arts art montréal canada travail université présenté travaille toronto festival	1.43%
Temporalité	temps eau corps plan moment silence mouvement action instant vitesse	1.42%
Marché de l'art	art valeur politique critique production économie marché artistes artistique conceptuel	1.05%
La textualité	mots texte livre voix textes langage lecture livres langue écriture	0.96%
Lumière	lumière nuit ciel couleur jour obscurité noire boîte blanc bleu	0.93%
Performance	performance performances corps action art actions public vidéo performatif performative	0.92%
Famille	mère père famille fille ans femmes vie enfants grand femme	0.80%
Rêve	rêve lit rêves jardin bête sommeil maïs yves proust mimétisme	0.64%
Jeu	sport jeu font goutte artaud siècle joueurs ball jeux carte	0.54%
Junk topic 1	bol ras texte queer instructions populaire maude inhospitalières punk short	0.44%
Films	film guerre films années fast cinéma paris événements politiques musée	0.40%
Junk topic 2	œuvres créatures œuvre terrariums forêt vidéo exposition spectateur ville artiste	0.37%
Voyages	camping tours tour sud visiteurs nord ouest pétrole lisa plaques	0.35%
Junk topic 3	favreau masques machine expe séripop christopher fournier ve szabo varady	0.35%

Tableau 1. Résultats de la modélisation thématique.

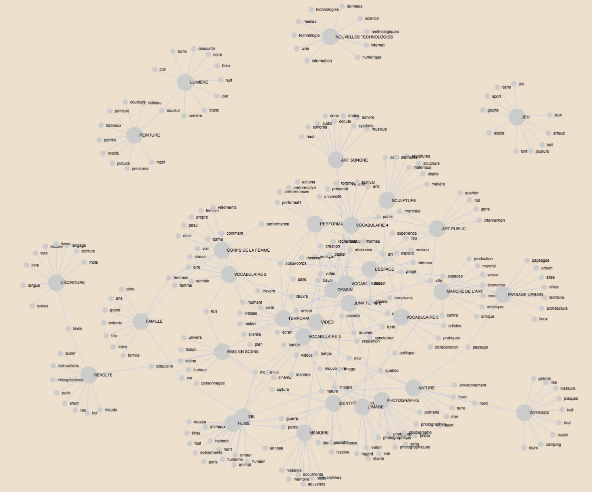


Figure 1. Résultats structurée de la modélisation thématique.

Une analyse des résultats obtenus met en lumière que la notion de thèmes se manifeste de différentes manière dans notre corpus. Ainsi, certains thèmes sont composés de mots-clés spécialisés et très caractéristiques des textes d'expositions (installation résidence création recherche temps espace, par exemple). Par ailleurs, d'autres regroupements thématiques renvoient sans équivoque à une discipline ou à une technique artistique (vidéo film images vidéos cinéma image bande œuvre). Les résultats du modèle thématique obtenu à partir de notre corpus de textes en art contemporain peuvent être synthétisés de la manière suivante :

1. les mots (les thèmes étiquetés « vocabulaire ») couramment utilisés (comme les verbes voir, faire, sentir) dans les textes d'exposition ou dans le discours artistique
2. les thèmes en art contemporain (révolte, famille, mémoire, rêve)
3. les disciplines ou techniques en art (peinture, sculpture, vidéo, photographie)
4. les thèmes incohérents (« junk topics ») (c'est-à-dire les regroupements sans cohérence apparente).

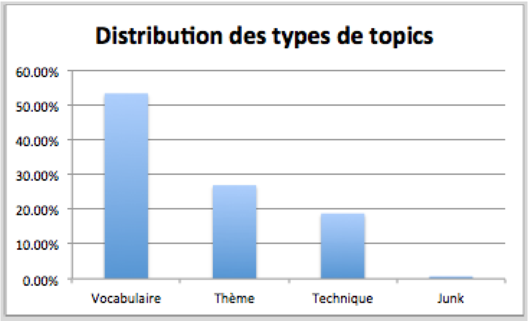


Figure 2. Distribution des types de thèmes.

Nous constatons la prépondérance des thèmes traitant du vocabulaire en art contemporain par rapport aux regroupements par sujets ou par technique (figure 2). En effet, la majorité des documents de notre corpus partagent un vocabulaire de base commun, un lexique propre aux textes d’exposition, constituant en quelque sorte le jargon du milieu. En excluant les mots vides, les textes d’exposition semblent donc composés surtout de mots sur le travail général de l’artiste en art (artiste, exposition, travail, projet, création, etc.). Le reste du corpus est consacré aux thématiques et à la technique.

Nous avons cherché à analyser les résultats du modèle thématique en diachronie afin d’observer l’évolution des thèmes dans le temps. À cette fin, nous n’avons conservé que les documents ayant été publié depuis les 16 dernières années, puisque nous ne disposions de très peu de documents dont la publication était antérieure à l’an 2 000. Nous présentons ici que certains thèmes pour lesquels une variation chronologique significative a été notée. Ainsi, les thématiques (figure 3) du “marché de l’art” et la “nature” ont fait l’objet d’un traitement croissant, alors qu’il en va inversement pour les thèmes du “rêve”, de la “famille”, de la “temporalité” et du “corps de la femme”.

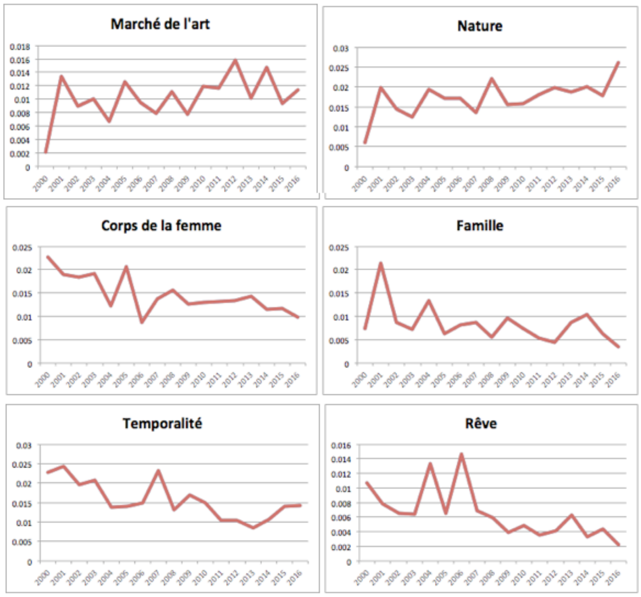


Figure 3. Thèmes à variation chronologique significative.

Les discours sur les techniques telles que la sculpture, l’art sonore ou les technologies numériques sont à la hausse depuis 2 000. Nous constatons que le discours sur la peinture est relativement stable, alors que celui sur la photographie et l’art public sont en baisse après avoir connu un intérêt marqué au début du 21e siècle (figure 4).

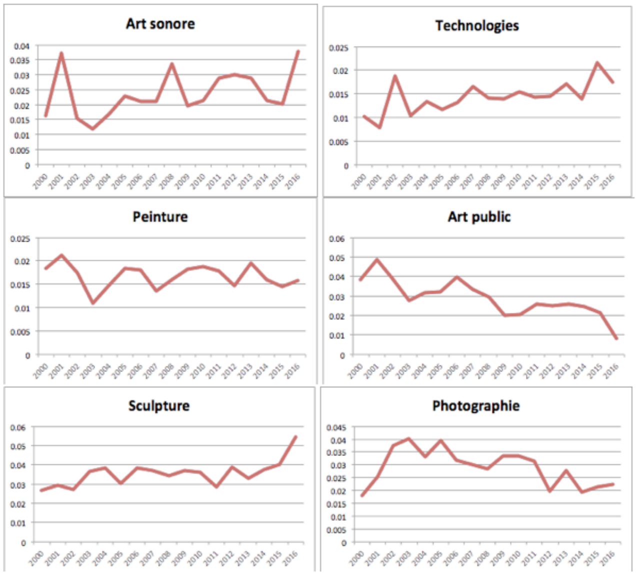


Figure 4. Les techniques à variation chronologique significative.

Conclusion

Dans le cadre de ce projet, nous avons appliqué la modélisation thématique sur un corpus de textes d’expositions en art contemporain dans les centres d’artistes autogérés du Québec. Les résultats témoignent de l’hétérogénéité des thématiques abordées. Ainsi, certaines thématiques concernent des techniques particulières, alors que d’autres reflètent le vocabulaire spécialisé en art contemporain. Une analyse diachronique des résultats nous a permis de constater l’évolution des thématiques abordées dans les textes produits par les centres d’artistes autogérés du Québec. Cette approche basée sur la notion de “distant reading”, appliqué en histoire de l’art, nous a permis d’observer d’un point de vue macro à la fois les tendances d’un milieu artistique et le vocabulaire communément utilisé dans les textes d’exposition.

Bibliography

Bender, K. (2015). Distant Viewing in Art History. A Case Study of Artistic Productivity. Journal of Digital Art History, (1), 101-109.

Blei, David M. (2012). Probabilist Topic Models. Communications of the ACM, 55(4), 77-84.

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, (3), 993-1022.

Blei, D. M. et Lafferty, J. D. Topic Models. In Text mining: classification, clustering, and applications. CRC Press, pp. 71-93.

Graham, S., Milligan, I., & Weingart, S. B. (2016). Exploring big historical data: The historian's macroscope. Imperial College Press.

McCallum, Andrew. (s.d.). MALLET homepage. Machine Learning for Language Toolkit. Repéré 22 avril 2016, à http://mallet.cs.umass.edu/

Réseau art actuel - Répertoire des membres. (s.d.). Réseau Art Actuel. Repéré 20 mai 2016, à http://www.rcaaq.org/html/fr/membres/index.php

Rhody, L. M. (2013). Topic Modeling and Figurative Language.

Journal of Digital Humanities. Repéré 21 mai 2016, à http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Triple Canopy – International Art English by Alix Rule & David Levine. (s.d.). Triple Canopy. Repéré 30 octobre 2016, à https://www.canopycanopycanopy.com/issues/16/contents/international_art_english

Enhanced Reflectance Transformation Imaging for Arts and Humanities

Peter Fornaro
peter.fornaro@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Andrea Bianco
andrea.bianco@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Aeneas Kaiser
aeneas.kaiser@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Lothar Schmitt
lschmitt.basel@gmail.com
Digital Humanities Lab, University of Basel
Switzerland

Heidrun Feldmann
heidrun.feldmann@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Introduction

The Digital Humanities Lab (DHLab) is a research group within the faculty of Humanities of the University of Basel. The research profile of the DHLab integrates computer science, digital imaging, computational photography and the accessibility of digital objects in humanities research. The project “Digital Materiality”, an interdisciplinary project in collaboration with the Seminar of Art History, examines how new digital methods and techniques can be used to describe the reflection of light on surfaces of artworks. Of main interest are mosaics and early prints; both categories have a strong interaction with light and standard photographic approaches are not able to capture the dynamic component of the light-surface interdependence that is specific for this kind of objects.

For art historians who study and work with mosaics or any other object of complex surface composition, it is difficult to capture and visualize the important surface features in such a way that research can be done with the reproduction. The major problem is the static nature of photographic reproductions, which does not allow interaction. Static, two-dimensional photographs cannot visualize appropriately the sparkling effect caused by the surface properties of the countless light reflecting tesserae of a mosaic, for example. Similar considerations apply for early prints, books and parchments. The visual impression that these objects convey is hardly delivered by a photograph. Metallic inclusions give the artwork a dynamic appearance caused by the change of reflectance behavior of glossy compared to matte material. Furthermore, the scholar may want to integrate information coming from other types of scientific photographs, such as infrared or ultraviolet illuminated or induced fluorescence images, to increase the visual impression of renderings of artwork. In particular, the combination of such scientific photographs with other imaging techniques like Reflectance Transformation Imaging is advantageous because multiple visual impressions can be combined in a way that would not be possible in reality.

Reflection Transformation Imaging

RTI (Reflectance Transformation Imaging) (Mudge et al, 2007; Malzbender et al, 2001) is a promising approach to go beyond the limitations of conventional photographic methods. However, RTI, as it is used today, has some drawbacks that are critical in the context of the reproduction of mosaics, metallic inclusions or most other artwork.

RTI makes it possible to interactively display the light reflection as a function of the light incident angle and the structure of the surface captured. RTI needs a series of digital images from a fixed camera position, in which the light source is illuminating the object in each capture from a different position. In order to produce RTI renderings of objects of different size, several lighting systems have been developed that consist of a number of light sources that are mounted on the inside of a hemispherical dome. Those light sources can be switched on and synchronized sequentially with a digital camera. The fixed design ensures a fixed and equal distance between the numerous light sources and the object. Thus, with such an automated setup the image capture process is drastically accelerated and, in addition, ensures a high reproducibility. In the Digital Humanities Lab such a dome has been developed as well; In this specific case it is an alloy design, which is equipped with 50 LEDs and an electronic control system, which enables the automatic sequential triggering of the light sources and the synchronization with the camera. The image sequence, recorded in this way, serves in a second step as data basis for a subsequent pixel-based modeling. For this purpose, a mathematical model - normally a function of second order - is fitted for each pixel position, which represents the set of all image points from the different directions of illumination, i.e., it is parameterized so that the square mean error of all data points relative to the function curve becomes minimal.

The reflection model so far used in the original RTI method described by Hewlett Packard (Mudge et al, 2006) corresponds to the mentioned simple second-order function. Thus it has only a relatively low mathematical complexity and therefore a limited precision to represent the actual surface reflection. This specific function, however, describes matt surfaces, a Lambertian radiator, very precisely. However, this method is not suitable for glossy materials, since this simple model does not correspond to the physical laws of gloss reflection. A further disadvantage is that the method, as it is usually used today, needs a specially designed viewer application.

In the Digital Humanities Lab, the method has been enhanced in order to be able to present more complex surfaces that are made from different materials. In our approach the complexity of the mathematical model is increased so that we can handle diffuse and gloss surface components at the same RTI image. In other words, this improvement makes it possible to model and display materials with different gloss levels using the same mathematical model.

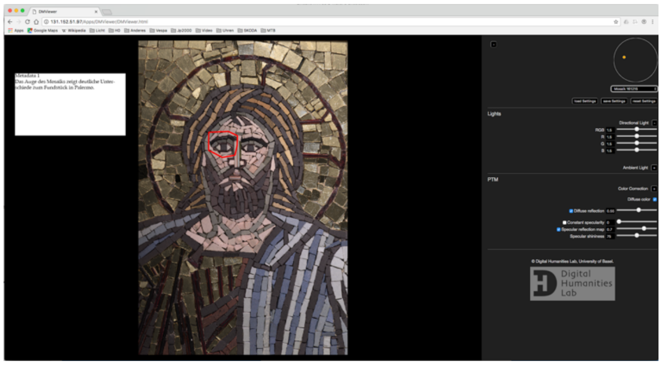
A suite for humanities scholars ‘needs

For humanities research another important aspect is the compatibility to web-based Virtual Research Environments (VRE). In principle, a functional graphical "front-end" with a connection to a database can be called a VRE. The aim of such a digital infrastructure for research is to allow scholars to work with methods and tools in the digital domain as they would do it in a conventional “analogue” process. This should be done in such a form that the scientists can intuitively recognize and use the well-known concepts and working methods that are offered by the VRE. This requires some basic functions:

- In any case, a browser-based client-server solution is preferable to a stand-alone application. In such a way collaborative work can be more easily achieved.
- In many humanities disciplines intensive work is being done with image material and objects, in which specific areas (region of interest, RoI) are often to be emphasized. Therefore corresponding graphic elements (lines, polygons) are necessary, with which such areas can be marked. These graphical elements can, for example, be polygonal line sections or rectangles, which allow the marking of object parts.
- The method of evaluating image material, which is common in the human sciences, is of a more qualitative nature. This kind of work requires extensive and powerful tools to capture descriptive and contextual metadata (annotations, transcriptions, comments). These meta-objects have to be linked with the actual primary object and need also to be stored in this way.

The visualization in a VRE must be multi-media. Besides text, image, sound and video, it is also of advantage to be able to display objects in a virtual 3D space.

To be able to integrate an RTI solutions into a web-environment to support real-time collaborative work needs specific web-technologies (Palma et al, 2010; MacDonald and Robson, 2010). The presented RTI viewer is fully web-compatible and it can be integrated in most browsers to support high-quality client-side visualization; the key technology that allows this integration is WebGL. This OpenGL-based programming interface, which has been optimized for "embedded systems", is nowadays integrated into any modern Web browsers. WebGL is a license-free standard developed to work seamless in conjunction with the programming language JavaScript. For the application in a browser this means that 3D functionalities are provided without the need to load any additional plug-ins. The performance and range of functions of WebGL are impressive. WebGL is also supported on most mobile devices, such as smartphones and tablets, which further increases the range of applications. The improved performance of WebGL, in contrast to many other graphics libraries, also allows for fluid interactive work. Long rendering times are left out and the visual latency is minimal. WebGL offers a variety of functionalities, ranging from simple grid models to complex animated, textured and illuminated surfaces. The fact that the use of graphic elements for marking RoIs is easily possible with this large range of functions is, of course, self-evident.



Metadata 1
Das Auge des Mosaiks zeigt deutliche Unterschiede zum Fundstück in Palermo.

131.152.51.97/Apps/DMViewer/DMViewer.html

Directional Light
RGB 1.6
R 1.6
G 1.6
B 1.6

Diffuse reflection 0.56
Constant specularity 0
Specular reflection map 0.7
Specular shininess 75

A Reflectance Transformation Imaging recording that is represented in a browser using WebGL. Around the left eye of the sacred head is a branding to be recognized, as well as a corresponding commentary, left in the picture window. The parameters of the viewing situation are shown on the right side, which are also stored in the data model. (Source: DHLab, University of Basel)

The integration of all those technologies and new developments allows us to present an improved solution to reproduce and render surfaces of different materials (matt and glossy) in a fully web-based environment implemented in JavaScript and WebGL, running on standard computers and mobile devices. In addition to RTI image processing, photographs in the UV and IR domain can be captured, displayed and superimposed with the same system to allow the user to compare the same region under different light condition. For flexibility, performance and data permanence aspects our RTI image server will follow the International Image Interoperability Framework (IIIF). IIIF defines a standardized URL syntax to serve digital images online in the field of cultural heritage and research. The region of interest, resolution, rotation and the file format of a requested image can be indicated on the URL. SIPI, the Simple Image Presentation Interface, developed by DHLab, provides an IIIF compliant image server which is ideally suited for our scope. Due to the fact, that the front-end is compatible with any standard web-browser, it can be integrated in a virtual research environment using Knora. Knora is a software framework, developed by DHLab, for storing, sharing, and working with primary sources and data in the humanities. Knora builds the fundament for the Swiss National Data Center for Research Data in Humanities that is operated by the DHLab. The source code is publicly available on Github at the address in reference.

This integrated environment allows researchers to interactively control the viewing and light conditions of e g. a digital mosaic rendering. Regions of interest can be chosen, annotated, saved, and shared with other scholars. The full set of contextual and technical metadata is stored and time stamped to be fully reloadable and citable back to any point of its history of changes.

Conclusions

The combination of the Enhanced-RTI and the Scientific Image Viewer (SIV) enables us to convey the impressions of these highly dynamic light-surface interactions and the information provided by IR and UV imaging e.g. to researchers who cannot visit the actual artwork in situ. The visual impression can be enriched by meta information that can shared with other scholars. The presented RTI based solution is also helpful to document the current condition of objects more accurately, e.g. before and after a restoration. Sustainability and simplicity of RTI image data is drastically improved by the use of a IIIF server. The presented infrastructure allows the strict separation of image data and meta information. As a result, any RTI image rendering is fully reproducible and therefore perfectly suited for digital archiving, following the requirements of performance, permanence and reliability.

Bibliography

Mudge, M., Ashley, M., and Schroer, C. (2007) A digital future for cultural heritage. Available at: http://culturalheritageimaging.org/What_We_Do/Publications/cipa2007/CIPA_2007.Pdf

Malzbender, T., Gelb, D., and Wolters, H. (2001) Polynomial texture maps. Proceedings of the 28th annual conference on Computer graphics and interactive techniques (pp. 519–528).

Mudge, M., Malzbender, T., Schroer, C., and Lum, M. (2006). New Reflection Transformation Imaging Methods for Rock Art and Multiple-Viewpoint Display. In The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST2006) (pp. 195–200).

Palma, G., Corsini, M., Cignoni, P., Scopigno, R., and Mudge, M. (2010) Dynamic Shading Enhancement for Reflection Trasformation Imaging. ACM Journal on Computing and Cultural Heritage, (pp. 6:1-6:20). (September 2010)

MacDonald, L., and Robson, S. (2010) Polynomial Texture Mapping and 3D Representations. International Archives of Photogrammetry, XXXVIII(8), (pp. 422–427).

PERiSCOPE: Personalized Color Profiles
A modern methodology for rendering colour in digital media displays

Peter Fornaro
peter.fornaro@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Sofia Georgakopoulou
s.georgakopoulou@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
Digital Humanities Lab, University of Basel
Switzerland

Colour and colour perception are of great relevance for many scientific fields and constitute important aspects of the aesthetic appearance of most artwork. In art history, visual arts, media sciences and photography, colour is an important attribute for assessing and reviewing most objects. Virtual Research Environments, for example, allow scholars to access, compare, and discuss digital renderings of paintings, photographs, video and motion pictures. The quality of colour rendering of these digital objects is crucial for most analytical work because colour is one of the most important features that is communicated. We propose to revolutionize the current colour rendering methodologies on state-of-the art displays and printers, for better “true” colour management of digital still images, video, or motion pictures.

Since the early years of colour science, the 1931 CIE Standard Observer has been a major reference for colour reproduction (Guild, 1925, 1932; Wright, 1929). The CIE Standard Observer defines how we can quantify colours that we see in the linear XYZ, or the perception-based Lab colour space. There are, however, several disadvantages to the system, arising mainly from the inherent assumptions made during the (very controlled) colour matching experiments used to derive the colour-matching functions (MacDonald, 2015). Most significantly, the original approach describes only the average of the experimental results, while those results are based on a small group of test individuals, in this case, 49 British men. Since the original experiment does not address the physiology of an individual observer, it has several qualitative limitations that result in colour reproduction with limited accuracy.

Since the original Standard Observer experiment, several studies have been performed aiming to improve the original colour matching functions (CMFs). Stiles and Burch, in 1959, repeated the experiment this time with an increased viewing field of 10 degrees, from the 2 degrees used for the 1931 results (Stiles & Burch, 1959). More recently, in the ‘90s, groups critically analyzed the 1939 and 1959 results to derive more precise CMFs. This is the case with Stockman et al in 1999 (Stockman, Sharpe, & Fach, 1999), who were able to propose new spectral sensitivity functions for the S-cones (the eye’s short-wavelength light receptor), both with their own experimental methods and by analyzing the existing functions. Meanwhile, individual differences in CMFs were estimated by North and Fairchild in 1993, this time by performing a small number of colour matching measurements and using a computational model to derive the CMFs (North & Fairchild, 1993a, 1993b). Investigations of individual variabilities in detecting colour among colour-normal individuals were performed in 2015 by Asano et al. and showed higher than expected interobserver variability (Asano, Fairchild, Blondé, & Morvan, 2015).

There are multiple studies that have been investigating the connection between colour perception and physiological (Abramov, Gordon, Feldman, & Chavarga, 2012a)(Abramov, Gordon, Feldman, & Chavarga, 2012b) and cultural (Collier, 1973; Merrifield, 1971)(Winawer et al., 2007)(Roberson, Davidoff, Davies, & Shapiro, 2006) differences, however the mechanisms that control the individual colour perception (i.e. the specific way that each person sees, understand and responds to colour) are not fully understood.

In this project, we plan to revisit the Standard Observer Experiment (SOE) and test its validity for different portions of the population. We will employ current technologies to develop and use a modern experimental setup based on LEDs and emphasize on these specific parameters of interest:

- The target colour in the full range of wavelengths in the visible range
- The geometry (size) of the colourfield
- The background the colour is surrounded by (limited to a specific set of colours)
- The ambient light conditions

The group of probands shall comprise of a large pool of individuals representing different ages, genders, social background and cultures. Our experimental setup will help to simplify the measurements of individual tristimulus functions for a more generalized application.

The original SOE has already shown a large variety of individual results, even though the number of probands was very small. Especially for the red primary, the observers have shown large differences in the required colour-mix to get a perceived colour match between the target colour and the mixing colours. With the experimental results we will confirm these differences in the colour matching experiment and move further to quantify the deviation both in wavelength and intensity of the individual tristimulus functions. If the principal shape of the three functions is comparable and the frequency shift is small relative to the full visible range, the experimental setup needed to measure the individual colour-matching functions can be drastically simplified.

Additionally, we pursue the development of a simplified experimental setup that will have a more generalized use. Previous colour-matching experiments using LEDs provide invaluable insight into the technical regarding optics requirements and electronics design (Morvan, Sarkar, Stauder, Blonde, & Kervec, 2011). The setup will consist of a device that can illuminate a test area with a red, blue and green primary of adjustable intensity. With such a mobile colour-matching device, the SOE can be simplified. The simplification of the experiment allows for observers to characterize their individual colour perception in a fast and uncomplicated way. The long-term goal is to develop automated single-observer colour measuring, which can be used to create personalized colour profiles (PERsonalIzed COlour ProfilEs PERiSCOPE), with which we will be able to uniquely calibrate the user’s viewing instruments, i.e. screens, printers etc. that are part of a conventional ICC colour management workflow. A simplified but straightforward experiment is the calculation and rendering of an image based on the scaled integral of spectral image data, regarded as spectral reflectance information for each pixel, S(λ, px, py) multiplied by the spectral power distribution of the illuminant I(λ) and e.g. the two extrema of the CIE's color matching functions $x_{min}(\lambda)$, $y_{min}(\lambda)$, $z_{min}(\lambda)$ as well as $x_{max}(\lambda)$, $y_{max}(\lambda)$, $z_{max}(\lambda)$. Such an experimental calculation results in two significantly different images, each representing a rendering based on different experimental data sets of individuals of the SOE. Such different renderings might have a significant effect on the perception of image data and it is a step towards the separation and equalisation of the various effects in the human visual system.

The results and the consequences of the project will be assessed together with experts of art history, media sciences and psychology. We believe that the implications of using personalized color profiles for image renderings will have a strong impact on various fields within the (digital) humanities, which will pave the way for new findings and better understanding. How is e.g. the same rendering of a colour photograph seen by two observers with different physiology? What are the effects of the perception of colour if the colour is adjusted to the physiology and described by of the two observers? How is an observer reacting, if the two observer renderings are shown to the same observer, showing the differences clearly? What effect has such an adjusted colour rendering to the description of eg. colour aesthetics? What is the consequence of seeing an image which is rendered for colour according to the data other person, in other words, if we see an artwork through the eyes of another person, eg. the artist itself? Those questions will be discussed in an interdisciplinary framework with other disciplines, especially art history and media sciences.

In terms of research in cultural heritage and art in humanities universities, museums, and other art institutes, our studies can change the way research is performed and will affect the results of this research, as the tools will become available to introduce the aspect of personal perception. Art is to a great extent a personal experience and institutes have to invent ways to make it objective in order to be able to discuss it in more general terms and draw conclusions about its effect in populations, cultures and times that may be foreign to the researchers today. The PERiSCOPE will adapt an artefact's colour on a digital display according to the viewer's profile. Thus it will allow scientists to truly see the same colour image at the same time and will help reduce the subjectivity of judging an artefact's appearance.

The DHLab also operates DaSCH (The National Data and Service Center for the Humanities), which includes a Virtual Research Environment in the humanities for sustainable collaborative work with digital sources. This web-based infrastructure can be ideally used for the discussion and collaborative evaluation of our colour results. It would make it possible for researchers to get a deeper understanding of each other's comments and critique in terms of colour. It would also enable viewers to see a work of art in the way the artist - in case he or she is still living - perceives it. Finally, if we are able to also categorize colour profiles to fit different population groups, the PERiSCOPE will also enable colour specialists (researchers or even industry stakeholders) to use colour in a way that would make the greatest impact to a specific part of the population.

Our research can also greatly impact research in psychology, in terms of personal perception and emotion, by providing new, more accessible methods for measuring and assessing responses to colour stimuli. Much of the knowledge we have about colour perception today comes from the field of psychology. For every colour-related project, researchers have to recreate some form of the SOE. With the PERiSCOPE mobile setup, we will be able to offer a standardised, easy-to-use system that will provide an efficient, reliable and reproducible methodology for these studies.

Bibliography

Abramov, I., Gordon, J., Feldman, O., & Chavarga, A. (2012a). Sex and vision II: color appearance of monochromatic lights. *Biology of Sex Differences*, *3*(1), 21.

Abramov, I., Gordon, J., Feldman, O., & Chavarga, A. (2012b). Sex & vision I: Spatio-temporal resolution. *Biology of Sex Differences*, *3*(1), 20.

Asano, Y., Fairchild, M. D., Blondé, L., & Morvan, P. (2015). Color matching experiment for highlighting interobserver variability. *Color Research and Application*, *41*(5), 530–539.

Collier, G. A. (1973). Reviewed Work: Basic Color Terms: Their Universality and Evolution by Brent Berlin, Paul Kay. *Linguistic Society of America*, *49*(1), 245–248.

Guild, J. (1925). A trichromatic colorimeter suitable for standardisation work. *Transactions of the Optical Society*, *27*(2), 106–129.

Guild, J. (1932). The Colorimetric Properties of the Spectrum. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *230*(681-693), 149–187.

MacDonald, L. W. (2015, April). *Realistic Visualisation of Cultural Heritage Objects* (Doctor of Philosophy). (P. S. Robson, Ed.). University College London.

Merrifield, W. R. (1971). Berlin Brent and Kay Paul, Basic color terms: their universality and evolution. Berkeley and Los Angeles: The University of California Press, 1969. Pp. xi + 178. *Journal of Linguistics*, *7*(02), 259–268.

Morvan, P., Sarkar, A., Stauder, J., Blonde, L., & Kervec, J. (2011). A handy calibrator for color vision of a human observer. In *2011 IEEE International Conference on Multimedia and Expo*. https://doi.org/10.1109/icme.2011.6012090

North, A. D., & Fairchild, M. D. (1993a). Measuring color-matching functions. Part I. *Color Research and Application*, *18*(3), 155–162.

North, A. D., & Fairchild, M. D. (1993b). Measuring color-matching functions. Part II. New data for assessing observer metamerism. *Color Research and Application*, *18*(3), 163–170.

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2006). Colour categories and category acquisition in Himba and English. In *Volume II. Psychological aspects* (pp. 159–172).

Stiles, W. S., & Burch, J. M. (1959). N.P.L. Colour-matching Investigation: Final Report (1958). *Optica Acta: International Journal of Optics*, *6*(1), 1–26.

Stockman, A., Sharpe, L. T., & Fach, C. (1999). The spectral sensitivity of the human short-wavelength sensitive cones derived from thresholds and color matches. *Vision Research*, *39*(17), 2901–2927.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7780–7785.

Wright, W. D. (1929). A re-determination of the trichromatic coefficients of the spectral colours. *Transactions of the Optical Society*, *30*(4), 141–164.

From Jane Austen's original *Pride and Prejudice* to a graded reader for L2 learners: a computational study of the processes of text simplification

Emily Franzini
efranzini@etrap.eu
eTRAP Research Group
University of Goettingen, Germany

Marco Büchler
mbuechler@etrap.eu
eTRAP Research Group
University of Goettingen, Germany

Introduction

Authentic text and graded reader

One of the objectives of second language (L2) learning is to be able to read and understand a variety of texts, from novels to newspaper articles, written in the language of interest. These texts written with a native audience in mind are commonly referred to as authentic texts or ”real life texts, not written for pedagogic purposes” (Wallace, 1992). Authentic texts, however, can present too many obstacles for L2 learners with too low a level of knowledge. The complex language structures and advanced vocabulary of these 'real' texts can have the unwanted effect of demotivating the reader (Richard, 2001). The gap between the learner's limited L2 knowledge and the fluency of authentic texts creates an ideal space for graded readers. Graded readers are ”simplified books written at varying levels of difficulty for second language learners” (Waring, 2012). Through graded readers original classic works can be adapted to match the learner's level of knowledge, thus providing the ideal tool to tackle 'real' themes, narratives and dialogues.

From authentic text to graded reader

One such graded reader is a newly adapted version of Jane Austen's *Pride and Prejudice* (edition of 1813) that one of the authors of this paper wrote (Franzini, 2016) as part of a collection for learners of English as a foreign language (EFL).

For authors, the process of adaptation of a text for a learning audience is complex. In order to simplify the text the author will necessarily have to make grammatical changes and lexical substitutions following vocabulary lists, shorten the text by cutting out entire paragraphs and events, and in some cases eliminate entire chapters and characters. Together with these changes, which can be defined as 'structural' because they are dictated by hard requirements of length and standardised level of difficulty, the author will also make a series of judgment calls at a sentence and word level. These changes, which are here defined as 'cognitive', include processes that are more intangible and that are a consequence of a native author's 'feeling' that the original text is too difficult for learners. These include elaborating, clarifying, providing context and motivation for unfamiliar information and non-explicit connections (Beck et al., 1991).

Research Objective

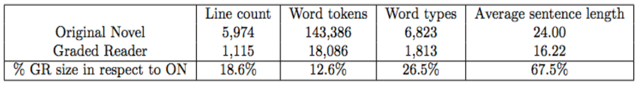
The objective of this study is to computationally analyse the manual process behind the simplification of a historical authentic text aimed at producing a graded reader. More specifically, it aims to classify and understand the structural and cognitive processes of adaptation that a human author, more or less consciously, is able to perform manually. Do the applied changes follow strict rules? Can they be classified as forming a pattern? And if so, can they be reproduced computationally?

Related Research

Researchers have long been addressing the issue of text simplification for a variety of purposes. A similar study to this was made by Petersen who compared authentic newspaper articles with abridged versions (Petersen and Ostendorf, 1991). Similar studies have been made, for example, to create a reading aid for people with disabilities (Canning, 2000, Allen, 2009).

Data

This study considers two sets of data. The first is the entire original novel (ON) *Pride and Prejudice*. The second dataset the graded reader (GR) published by Liberty. The GR has been compressed from the 61 chapters of the ON to 10 chapters. When comparing word tokens, the GR has a size of 12.6% of the ON (Tab. 1). The language was simplified to match the upper intermediate level B2. To guide the choice of vocabulary, the author chose to follow the Lexitronics Syllabus (Lexitronics, 2009).



	Line count	Word tokens	Word types	Average sentence length
Original Novel	5,974	143,386	6,823	24.00
Graded Reader	1,115	18,086	1,813	16.22
% GR size in respect to ON	18.6%	12.6%	26.5%	67.5%

Table 1: Quantitative comparison between data sets

Methodology

Readability

As a first step towards analysing the differences and similarities between an authentic text and a graded reader, we decided to evaluate if what is published as a graded reader can computationally be considered a simplified version of the original. The method chosen to make this investigation was to conduct two different readability tests, namely the ARI test and the Dale-Chall Index test on the data. Both tests were designed to gauge the comprehension difficulty of a text by providing a numeric value, which corresponds to a particular school level of a native speaker of the language tested.

The results show that both tests yield similar scores and satisfy the hypothesis that this particular GR can be computationally proven to be, in terms of 'understandability', a simplification of the ON.

	ARI	Dale-Chall
Original Novel	14-15 year olds	14-16 year olds
Graded Reader	11-12 year olds	11-13 year olds

Table 2: Age level of text understandability

Difference Analysis

In order to analyse the process of adaptation, a difference analysis was conducted by considering both those elements that changed from the ON to the GR, and those that, by contrast, remained the same. The analysis is structured into chapters, sentences and words, so as to proceed in order from the largest unit of text to the smallest.

When adapting a text, whether it is for a graded reader, a play or a film, the rationale behind the selection of certain parts over others is normally content-based. Here the author selected the most dynamic parts of the novel, which included dialogues, moments of suspense, movements of the characters and revelations. The selection of some scenes of the plot over others is purely a 'cognitive' choice of the author because it is entirely subjective. However, by using a text reuse detection software (TRACER) on both texts it was possible to visualise where the majority of reuses occur. These concentrate in particular around the beginning and the end of the novel (dark green in Fig. 1).

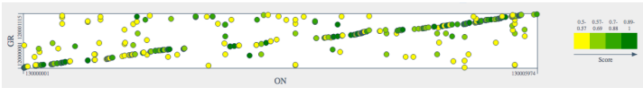


Figure 1: Dotplot visualisation of the reuses between the ON and the GR. The longer X-axis represents the larger original novel, the Y-axis the smaller GR. The darker the dot, the closer the similarities between the two datasets

'Structural' changes made at a sentence level present patterns that can be more systematically identified. For example, by comparing sentence length, it was noted that on average the ON contains longer sentences (24 words) than the GR (16.22 words) (Fig. 2). Though this might seem like an obvious result, it appears less so when one thinks that, in order to simplify a concept for a language learner, it is often necessary to use additional words to elaborate or clarify it.

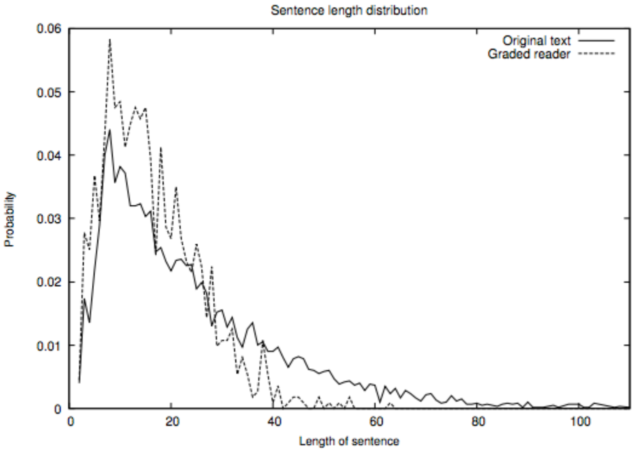


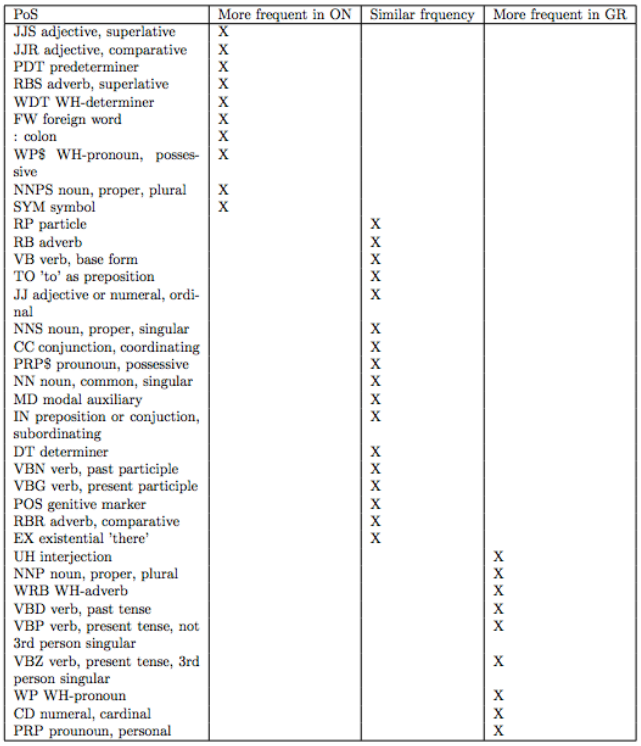
Figure 2: Sentence length distribution. The X-axis represents the number of words per sentence; the Y-axis is the probability of sentences of a specific length occurring in the texts

In order to conduct a difference analysis on the smallest unit of text - the word - we looked at all the words that appear frequently in the ON, but that never appear in the GR, in order to understand what kinds of words the author found necessary to drop.

Word	Frequency	Word	Frequency
upon	75	table	31
least	65	astonishment	30
acquaintance	63	fancy	30
either	59	attempt	29
whose	59	dine	29
dare	53	beg	28
regard	53	depend	28
determine	47	highly	28
scarcely	45	satisfaction	28
ladyship	42	acknowledge	27
former	38	credit	27
put	36	thus	27
amiable	35	disposition	26
deal	34	exceedingly	26
design	32	praise	26
satisfy	32	pray	26
society	32	wholly	26

Table 3: Words that appear only in the ON

Table (3) shows that 14 out of the 34 words listed (ca. 35%) are too advanced for level B2. Some of the other words, though accessible to B2 learners, were replaced with easier synonyms. We also conducted an analysis on parts of speech and how they differ in the two data sets (Tab. 4).



PoS	More frequent in ON	Similar frquency	More frequent in GR
JJS adjective, superlative	X		
JJR adjective, comparative	X		
PDT predeterminer	X		
RBS adverb, superlative	X		
WDT WH-determiner	X		
FW foreign word	X		
: colon	X		
WP$ WH-pronoun, possessive	X		
NNPS noun, proper, plural	X		
SYM symbol	X		
RP particle		X	
RB adverb		X	
VB verb, base form		X	
TO 'to' as preposition		X	
JJ adjective or numeral, ordinal		X	
NNS noun, proper, singular		X	
CC conjunction, coordinating		X	
PRP$ prounoun, possessive		X	
NN noun, common, singular		X	
MD modal auxiliary		X	
IN preposition or conjuction, subordinating		X	
DT determiner		X	
VBN verb, past participle		X	
VBG verb, present participle		X	
POS genitive marker		X	
RBR adverb, comparative		X	
EX existential 'there'		X	
UH interjection			X
NNP noun, proper, plural			X
WRB WH-adverb			X
VBD verb, past tense			X
VBP verb, present tense, not 3rd person singular			X
VBZ verb, present tense, 3rd person singular			X
WP WH-pronoun			X
CD numeral, cardinal			X
PRP prounoun, personal			X

Table 4: Parts of speech frequency in the ON vs. in the GR. Note the presence of comparative and superlative adjectives in the ON, which are totally absent from the GR

Conclusions and further research

This study is a first step into the realm of text simplification and adaptation regarding graded readers for L2 learners. By conducting a difference analysis between the two texts, it was observed that at plot level the selection of scenes has no impact on the difficulty of a text. The text reuse detection software used, however, identified which parts of the plot have been preserved and which have been eliminated for the sake of a consistent, yet shorter, story line. It was observed that the beginning and the end of the novel were the parts that were adapted most faithfully.

The identification of reuse over the whole novel was also a step towards pinpointing where sentences were reused verbatim and where they were not. Where the sentences have undergone heavy changes, we can observe to what extent they were modified, how and why. At a sentence level, we noted that reducing the length of the sentences is a successful simplification strategy. A further study would have to be conducted to best understand how sentences were split or reduced, and consequently how the syntax of a sentence was affected by its shortening.

At a word level, the simplification of the text appeared to be dictated by the elimination and replacement of difficult vocabulary and certain parts of speech, such as comparative and superlative adjectives. The word length does not appear to be an indicator of difficulty. While it was observed that both the readability tests were based on sentence length as a parameter, only the ARI test, however, considers word length as another parameter. A test on the word-length distribution of the ON versus the GR shows that, in this case, the word length bears no importance in assessing the difficulty of a text. Further research would have to be conducted in order to learn if it is easier for an L2 learner to remember a word not because of its length, but because of its repeated presence in a text. The insights gained from this study will be useful in future work on automating the simplification process.

Bibliography

Allen, D. (2009). "A study of the role of relative clauses in the simplification of news texts for learners of English." *System,* 37(4): 585–599.

Beck, I. L., McKeown, M. G., Sinatra, G. M., and Loxterman, J. A. (1991). "Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility." *Reading Research Quarterly*, Vol. 26(No. 3): 251–276.

Canning, Y. (2000). "Cohesive regeneration of syntactically simplified newspaper text." *Proc. ROMAND*, pp. 3–16. 14.

Council of Europe (n.d.) European CEFR - Common Framework of References for Languages. Language Policy of the Council of Europe: http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp

Franzini, E. (2016). *Adapted Edition of Jane Austen's Pride and Prejudice.* Liberty Publishing.

Lexitronics (2009). Lexitronics Syllabus. https://tvo.wikispaces.com/file/view/20386024-Common-English-Lexical-Framework.pdf

Petersen, S. E. and Ostendorf, M. (1991). "Text simplification for language learners: A corpus analysis." *Speech and Lanuguage Technology in Education SLaTE2007.*

Richard, J. C. (2001). *Curriculum development in language teaching.* Cambridge, C.U.P.

Wallace, C. (1992). *Reading.* Oxford, O.U.P.

Waring, R. (2012). "Writing graded readers." Accessible at: www.robwaring.org/papers/Writing_graded_reader.doc

Orosius' *Histories*: A Digital Intertextual Investigation into the First Christian History of Rome

Greta Franzini
gfranzini@etrap.eu
Georg-August-Universität Göttingen, Germany

Marco Büchler
mbuechler@etrap.eu
Georg-August-Universität Göttingen, Germany

Introduction

The research described in these pages is made possible by openly available Classical texts and linguistic resources.

It aims at performing semi-automatic analyses of Paulus Orosius' (385-420 AD) most celebrated work, the *Historia adversum Paganos Libri VII,* against its sources. The *Histories*, as this work is known in English, constitute the first history (752 BC to 417 AD) to have been written from a Christian perspective. To do so, Orosius drew from earlier and contemporary authors, including the pagans Caesar, Vergil, Suetonius, Livy, Lucan and Tacitus, thus providing a rich narrative fraught with intertextual references to poetry and prose alike.

Orosius' vast network of references challenges automatic text reuse detection tasks both qualitatively and quantitatively. In fact, information retrieval algorithms face differences in reuse style –from verbatim quotations to paraphrase and allusions (Navarro, 1991)– and millions of words to sift through. To understand how Orosius reused texts, it is necessary to detect, extract, classify and evaluate all references and compare them to their sources, mindful of the balance between the *precision* of the results and their *recall* or number.

Related Work

Existing research on Orosius' sources for the *Histories* is scattered and often focusses on his relation to one author or work only, albeit acknowledging the full spectrum of sources (e.g. Coffin, 1936; Sihler, 1887). The size of the source- texts (see Table 1) makes it extremely difficult to produce a comprehensive and detailed manual exploration of *all* of Orosius' references.

"It would be burdensome to list all of the Vergilian echoes [...]" (Coffin, 1936, p. 237)

What Coffin describes as "burdensome" can be accomplished with machine assistance. To the best of our knowledge, no existing study, traditional or computational, has quantified and analysed the reuse habits of Orosius.

The *Tesserae* project, which specialises in allusion detection, is the most similar to the research presented here (Coffee, 2013). The main difference between the two projects is that while *Tesserae* "aims to provide a flexible and robust web interface for exploring intertextual parallels", the present project aims at using intertextual parallels to optimise detection tasks and refine existing linguistic resources. The difference between the technologies is that our software gives users more control over the algorithmic process, offering a choice between being guided by default settings or supervising the detection by adjusting search parameters.

Research Questions and Contribution

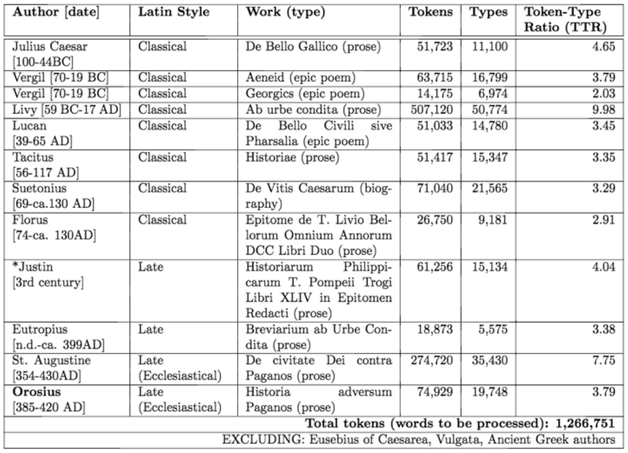
Our research began with the following questions: how does Orosius adapt Classical authors? Can we categorise his text reuse styles and what is the optimal precision-recall retrieval ratio on this large historical corpus? How does detection at scale affect performance?

In answering these questions, this project contributes new knowledge to both NLP and the Humanities. On the one hand, it serves as a case study for the testing of linguistic resources for Latin, and seeks to establish a workflow for the detection and evaluation of Latin text reuse at scale. On the other, it is the first attempt of its kind to analytically assess the reuse behaviour of Orosius and how his sources influenced his writing. This linguistic and literary analysis does not presume to subvert existing research on the topic but, rather, intends to facilitate philological enquiry by providing a detailed data-set from which inferences can be made, such as identifying unnamed sources or uncovering previously unnoticed reuses.

Data

All of the public-domain works for this study were downloaded from *The Latin Library*. We chose this collection over other analogous resources as it provides clean and plain texts (.txt), the format required by our text reuse detection machine TRACER.

Table 1 outlines the authors and works under investigation in chronological order. To give an idea of the size of the texts, the 'Tokens' column provides a total word-count for each work; the 'Types' column provides the total number of unique words; and the 'Token-Type Ratio' shows how often a type occurs in the text (e.g. a TTR of 3 indicates that for every type in a text there are three tokens. Generally, the higher the ratio the less linguistic variance in a text). This table reveals the language and challenges we should expect when detecting reuse. For instance, Caesar, Lucan and Tacitus share similar text lengths but Caesar has a higher TTR; this tells us that Caesar's text has less linguistic variety than Lucan and Tacitus. Conversely, if we look at Suetonius in comparison to Lucan and Tacitus, we notice a larger text but a similar TTR. This indicates a high linguistic variance in Suetonius' text, and one that can prove challenging for text reuse detection.



Author [date]	**Latin Style**	**Work (type)**	**Tokens**	**Types**	**Token-Type Ratio (TTR)**
Julius Caesar [100-44BC]	Classical	De Bello Gallico (prose)	51,723	11,100	4.65
Vergil [70-19 BC]	Classical	Aeneid (epic poem)	63,715	16,799	3.79
Vergil [70-19 BC]	Classical	Georgics (epic poem)	14,175	6,974	2.03
Livy [59 BC-17 AD]	Classical	Ab urbe condita (prose)	507,120	50,774	9.98
Lucan [39-65 AD]	Classical	De Bello Civili sive Pharsalia (epic poem)	51,033	14,780	3.45
Tacitus [56-117 AD]	Classical	Historiae (prose)	51,417	15,347	3.35
Suetonius [69-ca.130 AD]	Classical	De Vitis Caesarum (biography)	71,040	21,565	3.29
Florus [74-ca. 130AD]	Classical	Epitome de T. Livio Bellorum Omnium Annorum DCC Libri Duo (prose)	26,750	9,181	2.91
*Justin [3rd century]	Late	Historiarum Philippicarum T. Pompeii Trogi Libri XLIV in Epitomen Redacti (prose)	61,256	15,134	4.04
Eutropius [n.d.-ca. 399AD]	Late	Breviarium ab Urbe Condita (prose)	18,873	5,575	3.38
St. Augustine [354-430AD]	Late (Ecclesiastical)	De civitate Dei contra Paganos (prose)	274,720	35,430	7.75
Orosius [385-420 AD]	Late (Ecclesiastical)	Historia adversum Paganos (prose)	74,929	19,748	3.79
					Total tokens (words to be processed): 1,266,751
					EXCLUDING: Eusebius of Caesarea, Vulgata, Ancient Greek authors

Table 1. Overview of analysed texts. Excluded texts will be included in a second phase of the project. As we are still processing Justin we exclude him from the discussion for now.

Reuse Styles

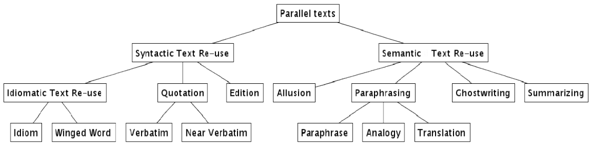


Figure 1. Graph illustrating reuse styles.

Orosius employs a variety of reuse styles, ranging from verbatim quotations to allusions and paraphrase (Navarro, 1991). The reuses are as short as two words (ibid.) or often invert the word order of the original text (Elerick, 1994).

Methodology

Our workflow makes use of three aids: a *TreeTagger* Latin language model for Part-of-Speech (PoS) tagging and lemmatisation (Schmid, 2013)– we chose to work with TreeTagger as, unlike other taggers, it comes with a pre-trained model for Latin (trained by Marco Passarotti); the *BabelNet 3.7* and *Latin WordNet* Latin lemma lists and synonym sets to support the detection of paraphrase and paradigmatically-replaced concepts; and TRACER, our text reuse detection machine**.**

First, the data is acquired and prepared: the texts are downloaded, cleaned through custom scripts and normalised. Next, the texts are lemmatised and tagged for PoS. Finally, we use TRACER to run detection tasks with different parameters in order to define the diversity of the reuses in the corpus.

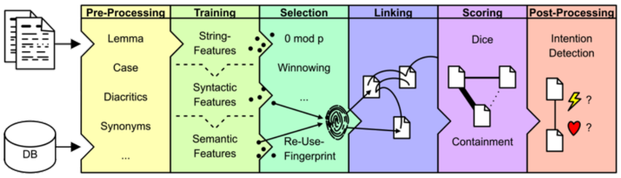
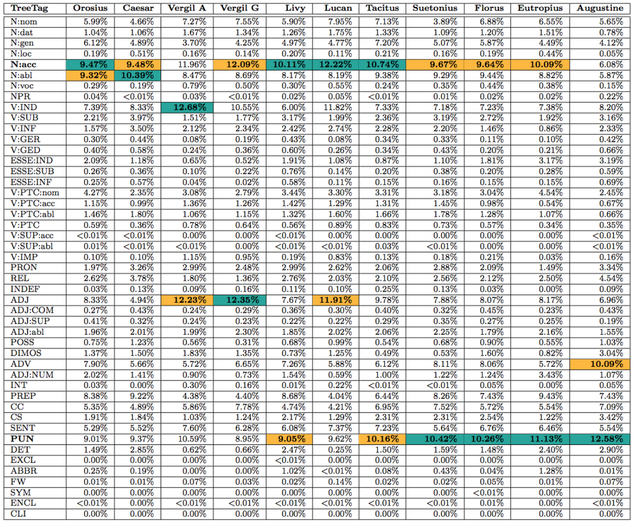


Figure 2: The six-step pipeline of TRACER (from left to right).

TRACER can split a detection task into six sub-tasks, each containing parameters that users can customise or (de)activate depending on the type of detection (see Figure 2). The reader will notice that the Pre-processing step also contains lemmatisation. This parameter is currently only available for English, hence our use of *TreeTagger* for Latin lemmatisation.

Results

Table 2 shows the distribution of *TreeTagger*'s PoS tags (in percentage) as rows across all authors analysed. Emerald-background cells indicate the most frequent PoS for an author (column), yellow-background cells the second-most frequent. Looking at the emerald cells, one can make two observations. First, we notice a higher use of punctuation during the shift from Classical to Late Latin, a result described in the full paper; second, we also notice the unsurprising predominance of the accusative case in most texts; accusative is the case of the object but is also used in expressions of time, place and space, in exclamations, after many prepositions and as the subject of the infinitive.



TreeTag	Orosius	Caesar	Vergil A	Vergil G	Livy	Lucan	Tacitus	Suetonius	Florus	Eutropius	Augustine
N:nom	5.99%	4.66%	7.27%	7.55%	5.90%	7.95%	7.13%	3.89%	6.88%	6.55%	5.65%
N:dat	1.04%	1.06%	1.67%	1.34%	1.26%	1.75%	1.33%	1.09%	1.20%	1.51%	0.78%
N:gen	6.12%	4.89%	3.70%	4.25%	4.97%	4.77%	7.20%	5.07%	5.87%	4.49%	4.12%
N:loc	0.19%	0.51%	0.16%	0.14%	0.20%	0.11%	0.21%	0.16%	0.19%	0.44%	0.05%
N:acc	**9.47%**	**9.48%**	11.96%	**12.09%**	**10.11%**	**12.22%**	**10.74%**	**9.67%**	**9.64%**	**10.09%**	6.08%
N:abl	**9.32%**	**10.39%**	8.47%	8.69%	8.17%	8.19%	9.38%	9.29%	9.44%	8.82%	5.87%
N:voc	0.29%	0.19%	0.79%	0.50%	0.30%	0.55%	0.24%	0.35%	0.44%	0.38%	0.15%
NPR	0.04%	<0.01%	0.03%	<0.01%	0.02%	0.05%	<0.01%	0.01%	0.02%	0.02%	0.22%
V:IND	7.39%	8.33%	**12.68%**	10.55%	6.00%	11.82%	7.33%	7.18%	7.23%	7.38%	8.20%
V:SUB	2.21%	3.97%	1.51%	1.77%	3.17%	1.99%	2.36%	3.19%	2.72%	1.92%	3.16%
V:INF	1.57%	3.50%	2.12%	2.34%	2.42%	2.74%	2.28%	2.20%	1.46%	0.86%	2.33%
V:GER	0.30%	0.44%	0.08%	0.19%	0.43%	0.08%	0.34%	0.33%	0.11%	0.10%	0.42%
V:GED	0.40%	0.58%	0.24%	0.36%	0.60%	0.26%	0.34%	0.43%	0.20%	0.21%	0.66%
ESSE:IND	2.09%	1.18%	0.65%	0.52%	1.91%	1.08%	0.87%	1.10%	1.81%	3.17%	3.19%
ESSE:SUB	0.26%	0.36%	0.10%	0.22%	0.76%	0.14%	0.20%	0.38%	0.20%	0.28%	0.59%
ESSE:INF	0.25%	0.57%	0.04%	0.02%	0.58%	0.11%	0.15%	0.16%	0.15%	0.15%	0.69%
V:PTC:nom	4.27%	2.35%	3.08%	2.79%	3.44%	3.30%	3.31%	3.18%	3.04%	4.54%	2.45%
V:PTC:acc	1.15%	0.99%	1.36%	1.26%	1.42%	1.29%	1.31%	1.45%	0.98%	0.54%	0.67%
V:PTC:abl	1.46%	1.80%	1.06%	1.15%	1.32%	1.60%	1.66%	1.78%	1.28%	1.07%	0.66%
V:PTC	0.59%	0.36%	0.78%	0.64%	0.56%	0.89%	0.83%	0.73%	0.57%	0.34%	0.35%
V:SUP:acc	<0.01%	<0.01%	0.00%	0.00%	<0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	<0.01%
V:SUP:abl	0.01%	<0.01%	<0.01%	0.00%	<0.01%	<0.01%	0.03%	<0.01%	0.00%	0.00%	<0.01%
V:IMP	0.10%	0.10%	1.15%	0.95%	0.19%	0.83%	0.13%	0.18%	0.21%	0.03%	0.16%
PRON	1.97%	3.26%	2.99%	2.48%	2.99%	2.62%	2.06%	2.88%	2.09%	1.49%	3.34%
REL	2.62%	3.78%	1.80%	1.36%	2.76%	2.03%	2.10%	2.56%	2.12%	2.50%	4.54%
INDEF	0.03%	0.13%	0.09%	0.16%	0.11%	0.10%	0.25%	0.13%	0.03%	0.00%	0.09%
ADJ	8.33%	4.94%	**12.23%**	**12.35%**	7.67%	**11.91%**	9.78%	7.88%	8.07%	8.17%	6.96%
ADJ:COM	0.27%	0.43%	0.24%	0.29%	0.36%	0.30%	0.40%	0.32%	0.45%	0.23%	0.43%
ADJ:SUP	0.41%	0.32%	0.24%	0.23%	0.22%	0.22%	0.29%	0.35%	0.27%	0.25%	0.19%
ADJ:abl	1.96%	2.01%	1.99%	2.30%	1.85%	2.02%	2.06%	2.25%	1.79%	2.16%	1.55%
POSS	0.75%	1.23%	0.56%	0.31%	0.68%	0.99%	0.54%	0.68%	0.90%	0.55%	1.03%
DIMOS	1.37%	1.50%	1.83%	1.35%	0.73%	1.25%	0.49%	0.53%	1.60%	0.82%	3.04%
ADV	7.90%	5.66%	5.72%	6.65%	7.26%	5.88%	6.12%	8.11%	8.06%	5.72%	**10.09%**
ADJ:NUM	2.02%	1.41%	0.90%	0.73%	1.54%	0.59%	1.00%	1.22%	1.24%	3.43%	1.07%
INT	0.03%	0.00%	0.30%	0.16%	0.01%	0.22%	<0.01%	<0.01%	0.05%	0.00%	0.05%
PREP	8.38%	9.22%	4.38%	4.40%	8.68%	4.04%	6.44%	8.26%	7.43%	9.43%	7.43%
CC	5.35%	4.89%	5.86%	7.78%	4.74%	4.21%	6.95%	7.52%	5.72%	5.54%	7.09%
CS	1.91%	1.84%	1.03%	1.24%	2.17%	1.29%	2.31%	2.31%	2.54%	1.22%	3.42%
SENT	5.29%	5.52%	7.60%	6.28%	6.08%	7.37%	7.23%	5.64%	6.76%	6.46%	5.54%
PUN	9.01%	9.37%	10.59%	8.95%	**9.05%**	9.62%	**10.16%**	**10.42%**	**10.26%**	**11.13%**	**12.58%**
DET	1.49%	2.85%	0.62%	0.66%	2.47%	0.25%	1.50%	1.59%	1.48%	2.40%	2.90%
EXCL	0.00%	0.00%	0.00%	0.00%	<0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ABBR	0.25%	0.19%	0.00%	0.00%	1.02%	<0.01%	0.08%	0.43%	0.04%	1.28%	0.01%
FW	0.01%	0.01%	0.07%	0.03%	0.02%	0.14%	0.02%	0.02%	0.05%	0.01%	0.07%
SYM	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	<0.01%	0.00%	0.00%
ENCL	<0.01%	0.00%	<0.01%	<0.01%	<0.01%	<0.01%	<0.01%	<0.01%	0.01%	0.00%	<0.01%
CLI	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 2. Part-of-Speech distribution across all analysed authors.

Tagged and lemmatised text accounts for 93.1% of the corpus. A 7% of words could not be lemmatised due to typos in the text (e.g. missing white-spaces), which we will correct; similarly, some words could not be successfully tagged, as the lemmas are not included in the parameter file.

To run TRACER, the texts were initially segmentised by sentence. The average sentence length measured across the whole corpus is 31 words per sentence. A first experiment at the sentence-level failed due to the presence of very short reuses. For this reason, the segmentation was changed to a moving window segmentation with a window-size of ten, restricting TRACER's detection from entire sentences to smaller windows. In the *Selection* step (see Figure 2), we experimented with max-pruning and a PoS selection. However, max-pruning did not perform well because the sentences were too long compared to the scored overlap (see Figure 3). TRACER produced precise results but the recall was too low. For this reason, we changed to a PoS-based selection, which considered nouns, verbs and adjectives as relevant features, thus significantly increasing the recall and the overall quality of the results.

For the *Scoring* (see Figure 2), we used the resemblance score, which measures the ratio of overlapping features with the overall unique set of features of two alignment candidates. Figure 3 illustrates the results of the research thus far: we notice that just over 50% of all scored alignment pairs of two text passages have a four-word overlap only (e.g. nouns, verbs, etc.), and that 93.9% of all candidates have overlaps of 3, 4 or 5 words, indicating a fragmentary reuse style rather than block-copying.

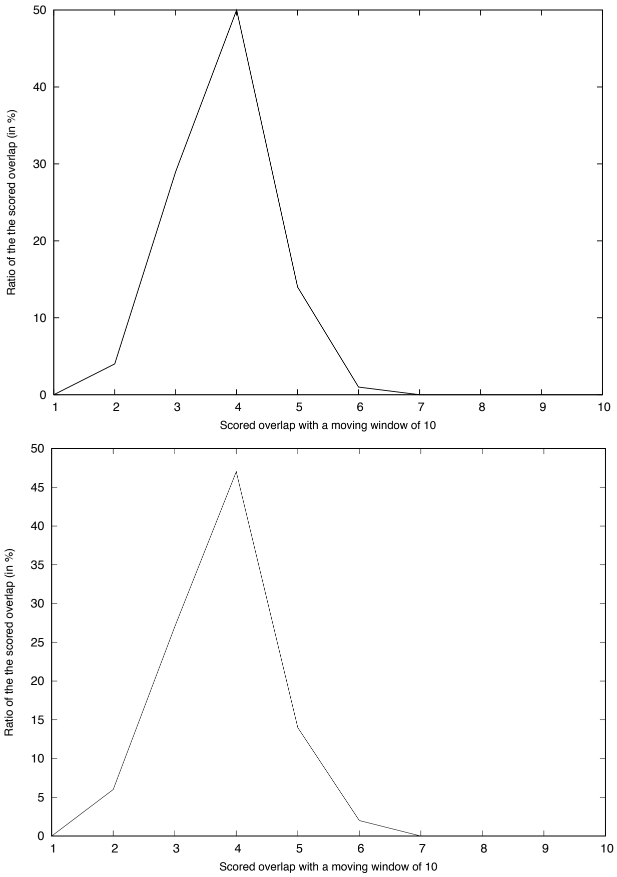


Figure 3. In these plots the x-axis represents a window of 10 words, while the y-axis the occurrence of the overlap in percentage. The left plot presents results identified across all texts; the right plot represents reuses identified in Orosius.

Orosius' plot in Figure 3 shows that over 45% of reuses identified overlap with source texts by 4 words, fitting the pattern identified across *all texts* displayed in the left plot of Figure 3.

Further Results

In the full paper, we describe the detection processes and how the computed results and known reuses correlate. We discuss the identification of paradigmatic relations against *BabelNet* and the *Latin WordNe*t for the detection of paraphrase and looser forms of reuse, and address the bridge between close and distant reading by providing a qualitative and quantitative study of Orosius' intertextual network. This includes a discussion on the reuse style of the author and a list of all reuses, both known and previously unnoticed by scholars. Furthermore, we describe how this research can contribute to the establishment of a Gold Standard for Latin lemmatisation and text reuse detection.

Limitations and Conclusion

This project aims at testing the stability of historical text reuse detection at scale, while analytically assessing the narrative and text reuse techniques of Orosius, our case study. We chose Orosius' *Histories* for its rich and diverse intertextual network, which tests the full spectrum of text reuse algorithms belonging to TRACER, giving us the opportunity to refine them against reuses identified in existing commentaries to Orosius, and to look at the whole array of reuses for further philological study into the degrees of influence his sources exerted upon his writing.

The retrieval accuracy of TRACER partly depends on the accuracy of the trained models of *TreeTagger* and on the *BabelNet* and *Latin WordNet* data. An error analysis is needed in order to verify the accuracy of our cleaned and automatically-tagged data, and to determine the effect of this incorrect tagging on text reuse detection. Depending on the outcome of this analysis, we will communicate the unlemmatised words to *TreeTagger* developers and/or consider re-tagging our corpus with a more advanced tagger, such as *LemLat* (Passarotti, 2007), incorporating curated linguistic resources (e.g. *Word Formation Latin* project) or even training a tagger on the different Latin types constituting our corpus.

Acknowledgements

This research is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

Bibliography

Büchler, M., Franzini, G., Franzini, E. Moritz, M. (2017 forthcoming). "TRACER - a multilevel framework for historical Text Reuse detection." *Journal of Data Mining and Digital Humanities - Special Issue on Computer Aided Processing of Intertextuality in Ancient Languages*.

Coffee, N., Koenig, J. P., Poornima, S., Forstall, C. W., Ossewaarde, R., Jacobson, S. L. (2013). "The Tesserae Project: intertextual analysis of Latin poetry." *Literary and Linguistic Computing,* 28(2): 221–28. DOI: 10.1093/llc/fqs033

Coffin, H. C. (1936). "Vergil and Orosius." *The Classical Journal*, 31(4): 235-41. Available at: http://www.jstor.org/stable/3290976 (Accessed: 13 October 2016)

Elerick, C. (1994). "How Latin Word Order Works." *Journal of Latin Linguistics*, 4(1): 99-118. DOI: 10.1515/joll.1994.4.1.99

Fear, T. A. (2010). *Orosius: Seven Books of History against the Pagans.* Liverpool University Press.

Navarro, M.A.R. (1991). "Historiadores y poetas citados en las Historias de Orosio: Livio y Tácito, Virgilio y Lucano." *Fortunatae: Revista canaria de filología, cultura y humanidades clásicas*, (2): 277-86. Available at: https://dialnet.unirioja.es/descarga/articulo/163829.pdf (Accessed: 13 October 2016).

Passarotti, M. (2007). "LEMLAT. Uno Strumento per la Lemmatizzazione Morfologica Automatica del Latino." *Journal of Latin Linguistics,* 9(3): 107–128. DOI: 10.1515/joll.2007.9.3.107

Schmid, H. (2013). "Probabilistic Part-of-Speech Tagging Using Decision Trees." In D. B. Jones and H. Somers (eds), *New Methods in Language Processing.* London and New York: Routledge, pp. 154-64. Available at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf (Accessed: 28 October 2016).

Sihler, E. (1887). "The Tradition of Caesar's Gallic Wars from Cicero to Orosius." *Transactions of the American Philological Association (1869-1896),* 18: 19-29. Available at:

http://www.jstor.org/stable/2935772 (Accessed: 14 October 2016).

Collaborations in the Global Midwest: The Diffusion of DH Values in Research Collaborations in the Humanities Without Walls Consortium

Harriett Green
green19@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Megan Senseney
mfsense2@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Maria Bonn
mbonn@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Introduction

The advent of new funding streams and initiatives within broader humanities scholarship indicate that the collaborative research approaches have diffused beyond digital humanities. This paper presents the findings of the "Humanities Collaboration and Research Practices: Exploring Scholarship in the Global Midwest" project (HCRP), which examines the Humanities Without Walls initiative as a case study for how innovative, interdisciplinary humanities research draws upon models from digital humanities.

Background

The Humanities Without Walls (HWW) Global Midwest initiative supports collaborative research projects led by faculty from fifteen U.S. research universities in the Midwest. With its emphasis on multi-institutional, interdisciplinary collaboration and applied research, HWW Global Midwest presents rich research cases on the evolving nature of humanities research.

Literature

Studies of collaboration among digital humanities researchers and its impact on humanities scholarship have proliferated over the past decade (Siemens, 2009; Siemens, 2011; Deegan & McCarty, 2012; Given & Wilson, 2015). Focused studies of DH research practices also examine credit and authorship (Nowviskie, 2011; Nowviskie, 2012), infrastructure needs (ACLS, 2006, Edmond, 2015), and project management (Leon, 2011). Building upon this research, our study examines how the collaborative experimentations undertaken by HWW Global Midwest researchers influenced their research practices, data sharing, and final outputs.

Method

The project team conducted semi-structured interviews with 28 researchers funded by the first round of HWW Global Midwest awards. Participants were asked about project goals, collaboration development, tools used for project management, challenges, and research approaches.

The project team recorded and transcribed the interviews, and coded them in ATLAS.ti 7. Each transcription was coded multiple times for inter-coder reliability. This study applies a qualitative content analysis method that expands upon prior studies by Brockman et al. (2001), Palmer & Neumann (2002), and Palmer (2005), as well as a grounded theory approach (Corbin & Strauss, 2008).

Findings

Interviews with Global Midwest project awardees revealed a number of emerging practices and challenges common to the DH community and collaborative DH projects.

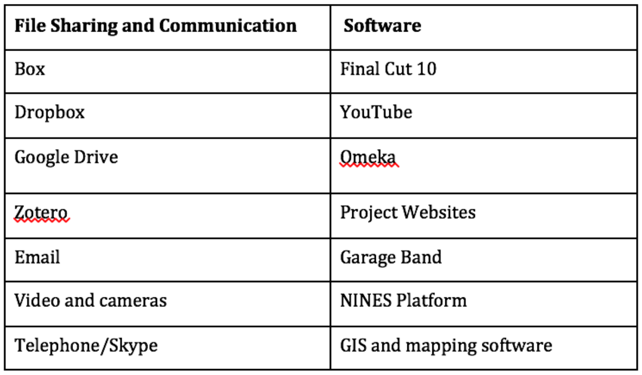
Project Workflows and Infrastructure

The interviewed participants identified many project challenges within the HWW program model. These included finding personnel and eligible collaborators, aligning IRB approvals, and funding coordination.

One participant summed up the sentiments of many on project management, saying "that was definitely a learning curve for all of us." But most deemed this learning curve worth undertaking.

Another key aspect of project workflows was the range of tools used by the HWW Global Midwest research groups (See Table 1). Tool selections ranged from cloud storage to unique platforms, including the software built for NINES and 18th Connect. But whether they used popular or specialized tools, one respondent's declaration captures their prevailing ethos: "We're using an existing infrastructure and we're applying it in a quite different way."

This process of translating tools to different uses is similar to the software adaptations seen in digital humanities research, and as scholars explore new ways to translate their research, they turn to multiple sources of expertise.



File Sharing and Communication	Software
Box	Final Cut 10
Dropbox	YouTube
Google Drive	Omeka
Zotero	Project Websites
Email	Garage Band
Video and cameras	NINES Platform
Telephone/Skype	GIS and mapping software

Table 1: Tools for Research

Methods of Collaborative Analysis

Many research groups carefully developed methods of analysis in ways that resonate with cross-disciplinary approaches in digital humanities research. One respondent characterized a group's work as having "a lot of cross-fertilization of methodologies ... not so much about content." Another project planned to employ several methods of analysis, including a short film, a series of interviews, and a performance of dancers and scholars rolling around on the floor "because to resist was not going to happen." This type of collaborative process was described by one group as one that "unfolds in an uncertain and, in that sense, an egalitarian manner because no one knows yet what the thing will be…. You go on a hunch and you see where it takes you. That is typical of ethnography, but also, I think, of collaboration, as well." These dynamic and educational elements of collaboration proved to be key to partnerships.

Student Engagement

Several interviews related a need for research assistance and dedicated project management, and respondents repeatedly attested to the value of graduate assistants who shouldered the management burden of the projects, or the (unfulfilled) need for such students. Projects navigated the tension between relying on student labor and acknowledging the intellectual contributions of the students with varying degrees of success, with the most positive assessment citing student participation as the true catalyst for collaborative practice: "They're not just graduate students. They're fellow collaborators in the project at this point and they have tremendous resources of knowledge, you know. The multiplication is enormous. It's here that you really have the collaborating humanities."

Digital Dissemination and Curation

Respondents cited different formats for sharing their work, including performances, films, and websites as well as texts and presentations. Several respondents envisioned creating hybrid outputs, such as one respondent's plan "to create some kind of interactive map [and] ideally a repository of sounds." Another discussed the possibility of sharing interview data as a form of dissemination, noting that "we're still processing the data [and] deciding how to feature it… we're not tweeting the results or something like that." This response also highlights the complex characteristics of humanities data, and the multiplicity of factors that must be considered for data sharing and archiving.

Respondents also saw avenues for making broader impacts via use of different platforms. As one respondent explained, "I think we've contemplated scholarly output in the traditional platforms… whether they're online or in print, but we have contemplated getting research into the hands of stakeholders who are not scholars."

Collaboration and Credit

Many respondents were mindful of the importance of providing appropriate credit and recognition for project partners. One respondent noted that "for us, the notion of collaboration was built around the idea that both parties would be equally acknowledged." Negotiating appropriate credit, however, also can reveal moments of tension within projects. Another respondent observed that "there was a little bit of misunderstanding, and some disagreements […] had to do with who is being acknowledged for what."

Respondents differed on their views of co-authored publications. One respondent noted, "I didn't expect a lot of co-authoring, more of a co-design of the platform." Another viewed co-authorship as an important "end product collaboration." While discussion of evaluation for tenure and promotion were present within the interviews, they were not as prevalent as might be expected. Yet a key theme that emerged in the responses was that culture shifts within humanities disciplines are essential to advancing the acceptance of research collaborations and co-authorship in peer evaluation criteria.

Discussion and Conclusion

To bring emergent humanities research collaborations into dialogue with the digital humanities, we propose a set of recommendations as a foundation for fostering rigorous interdisciplinary collaboration:

Build stronger connections between teaching and research through engaging students in research collaborations: Student participation in digital humanities projects has been essential to the growth of DH research, and humanities scholars can similarly bring collaborative research practice into the classroom in ways that acknowledge and recognize the students' labor.

Experiment with new forms of dissemination that more accurately convey the full breadth of collaborative work: HWW Global Midwest researchers frequently sought new ways for disseminating interdisciplinary research findings: In the same way that digital humanities researchers employ new formats for publishing research data and findings, humanities scholars can experiment with new forms that reflect interdisciplinary approaches. Scholars should also consider protocols for establishing credit and co-authorship, such as a negotiated project charter that establishes workflows for the collaboration, standards for co-authorship and a grievance process.

Encourage a culture of sharing data and interim findings: Administrators are in a key position to encourage shifts in humanities research practices by encouraging and explicitly ascribing value to related intellectual activities. Both leaders as well as researchers can encourage a culture of sharing data and interim phase research outputs that recognize the complexities of the communities and types of data in humanities research.

Strategically expand institutional investments in humanities research collaborations in order to ensure research sustainability: To ensure sustainable collaborations, administrators may need to make financial and structural investments, and key to these decisions is understanding the motivations and requirements of multiple stakeholder groups represented within a project. For example, some team members may require explicit funding to dedicate allocations of their time, while other team members may need support staff assistance to manage budgets and project documentation. Another avenue is to leverage the embedded collaborative power of regional, national, and international consortia in order to ensure research sustainability.

These recommendations drawn from our findings suggest that the expansion and sustainability of innovative research collaborations in the humanities has critical intersections with the evolving research practices of digital humanities research.

Bibliography

American Council of Learned Societies (2006). *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. Washington, DC: American Council of Learned Societies. http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf

Brockman, W. S., Neumann, L., Palmer, C. L., & Tidline, T. J. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment*. Council on Library and Information Resources. Retrieved from http://eric.ed.gov/?id=ED459855

Corbin, J. & Strauss, A. (2008). *Basics of qualitative research*. 3rd ed. Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781452230153

Deegan, M., & McCarty, W. (2012). *Collaborative Research in the Digital Humanities: A Volume in Honour of Harold Short, on the Occasion of His 65th Birthday and His Retirement, September 2010*. Ashgate Publishing, Ltd.

Edmond, J. (2015). Collaboration and Infrastructure. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A New Companion to Digital Humanities* John Wiley & Sons, Ltd., pp. 54–65. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118680605.ch4/summary

Given, L. M., & Wilson, R. (2015). Collaboration, Information Seeking, and Technology Use: A Critical Examination of Humanities Scholars' Research Practices. In P. Hansen, C. Shah, & C.-P. Klas (Eds.), *Collaborative Information Seeking*. Springer International Publishing, pp. 139–64. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-18988-8_8

Leon, S.M. (2011). Project Management for Humanists: Preparing Future Primary Investigators. In *#Alt- Academy, Volume One*. MediaCommons. http://mediacommons.futureofthebook.org/alt-ac/pieces/project-management-humanists (accessed September 23, 2016)

Nowviskie, B. (2011). Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. *Profession 2011*, 1: 169–81. https://doi.org/10.1632/prof.2011.2011.1.169

Palmer, C. L., & Neumann, L. J. (2002). The Information Work of Interdisciplinary Humanities Scholars: Exploration and Translation. *The Library Quarterly: Information, Community, Policy*, 72: 85–117.

Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *Journal for the American Society of Information Science*, 56: 1140-53.

Siemens, L. (2009). It's a team if you use 'reply all': An exploration of research teams in digital humanities environments. *Literary and Linguistic Computing*, 24: 225-33. https://doi.org/10.1093/llc/fqp009

Siemens, L. (2011). The Balance between On-line and In-person Interactions: Methods for the Development of Digital Humanities Collaboration. *Digital Studies / Le Champ Numérique*, 2: Retrieved from https://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/184

Indigenizing the Digital Humanities: Challenges, Questions, and Research Opportunities

Jennifer Guiliano
jenguiliano@gmail.com
Indiana University-Purdue University Indianapolis
United States of America

Carolyn Heitman
heitman@unl.edu
University of Nebraska-Lincoln
United States of America

In 2010, 2.9 million Americans self-identified as American Indian or Alaska Native. In addition, another 2.3 million people reported as American Indian or Alaska Native in combination with one or more other races. There are 562 federally-recognized tribes in the United States with dozens of other state-recognized tribes awaiting federal recognition. Outside of the U.S. there are millions of other indigenous community members in lands far-flung. Native American Studies (on Native American Studies and its relationship to knowledge, see Fixico, 2013; Madsen, 2012; Meyers, 2016; and Warrior, 2014), an interdisciplinary field of study exploring the history, culture, politics, issues, and contemporary experience of these indigenous peoples of America, intersects with a number of issues related to access, preservation, and methodology that are problematized through the development and deployment of digital tools, methods and research. While tremendous work has been done around the preservation and access of analog materials within Native American communities (for an example of digital Native American projects, see the Chaco Canyon Research Archive), there has been much less attention paid to the ways in which digital objects (on analog to digital surrogates, see the *Protocols for Native American Archival Materials*; on digital repatriation, see Christen, 2011; Colwell-Chanthaphonh, 2011; Colwell, 2015; and Runde, 2010), practices, and methods function within Native communities and through Native American Studies scholarship. With the exception of the Murkutu content management system which originated with the Warumungu Aboriginal community in the Central Australian town of Tennant Creek, digital humanities researchers and developers have been largely separate from Native American Studies– for example, it is only in the last year or so that we have begun to see digital humanities represented at Native American Studies conferences (e.g. Ethnohistory, Native American and Indigenous Studies Association, etc.) or within topical sub-panels at professional gatherings like the American Studies Association, the Modern Language Association, or the American Historical Association. There are hundreds of scholars actively working at dozens of American Indian, Native American, and Indigenous studies programs in addition to scholars working in humanities departments like literature, history, anthropology, archeology, etc. While museum, library, and archival communities have digitized Native cultural objects, literature, and music, there has been just cursory use of these materials within the larger digital humanities research landscape outside of digital archives or digital anthropology/archeology contexts. This submission opens by exploring a handful of individual projects that have leveraged Native materials to highlight how Native American studies scholars might critique the use and contextualization of tribal materials. Potential projects for exploration include Performing Archive: Edward Curtis + "the vanishing race" , the Indigenous Digital Archive, and the Native American Images Project. The digital humanities community has done little to inform itself of the unique issues associated with research and teaching by and in the Native American context. This knowledge has not been integrated for five reasons: 1) funding for Native American studies research institutionally has focused on the production of traditional scholarly products (e.g. monographs); 2) Native American studies as a discipline has remained underfunded institutionally which bars many institutions from hiring staff, faculty, and tribal members to work together to teach and research; 3) funding for cultural heritage organizations including tribal archives, libraries, and museums, has focused on preservation and access to analog materials (on intellectual property issues associated with analog to digital Native materials, see Anderson, 2005; and Brown and Nicholas, 2012) only, rather than embracing the spectrum of digital and analog materials that document Native life; 4) the diversity of academic disciplines who participate in Native Studies is broad and, as such, lends itself to fragmentation and a lack of information exchange about digital projects, tools, methods, and pedagogy; and 5) Digital Native studies most frequently is only encountered by scholars and community members at the end of the development cycle as the project is released and they begin to encounter "actual" users. As a result, there do not exist any best practices, guidelines, or even suggestions about the process of working in digital environments with Native American communities. Archaeologists and museum professionals have made the furthest strides in attempting to address engagement with tribal communities and research (on issues of information architecture and archiving in indigenous contexts, see Powell and Aiken, 2010; Cushman, 2013; Senier, 2014; Christen, 2012; and Joffrion and Fernández, 2015), yet that work has not widely proliferated among the interdisciplinary humanities nor the digital humanities more generally. As a result, each digital humanist (and humanist) must start from scratch and negotiate every encounter without guidance or lessons learned from those who have attempted integration of the digital with Native American Studies before.

Additionally, because these conversations have been focused mainly on individual projects or on knowledge management generally (e.g. the Murkutu Project and the Intellectual Property Issues in Cultural Heritage Protocol), the cross-pollination of an interdisciplinary community of digital researchers with tribal communities, cultural heritage organizations, and academic scholars has been significantly underdeveloped. Thus, it is not surprising that there are only a handful of funded digital projects documenting Native American life: the American Indian Treaties Portal, a digital collection of the final texts of 366 of the 375 American Indian treaties recognized by the United States Department of State, and the digitized Journals of Lewis & Clark Expedition created in partnership with the Center for Digital Research at the University of Nebraska Lincoln. Mark of the Mississippians: A Multi-Platform Digital Media Project (Cahokia Mounds Museum Society) and Meeting the Earthworks Builders, a flashbased video game are currently being developed by the Ohio State University.. Additionally, NEH Public Programs just funded Indians of the Midwest, an educational website focused on recent scholarship on Native peoples and the Newberry Library Collection.

The barriers to digital fluency in Native American studies are varied and include such obstacles as cultural rules regarding access to sensitive materials, the advanced technical expertise that software and hardware often requires beyond basic digitization, the costs of digital infrastructure and proprietary license fees, and issues of community engagement and trust that might limit the display of digital materials about Native peoples. It might be tempting to assume that the uptake of digital humanities method and pedagogy in the academic, cultural heritage, and tribal communities is one of lack of information or funding, but in fact the digital humanities researchers included as part of the Digital Native American and Indigenous Studies project are finding that it is also the cultural barriers to access, display, and analysis across differing types of digital materials that are challenging our ability to leverage digital tools, resources, and approaches.

Digital Humanities articulates three parallel interdisciplinary commitments to “openness”: 1) a commitment to open access publishing; 2) a commitment to open access/open source software development; and 3) a commitment to open access data. While the first two trends have received deep and lasting attention via scholarly publishing and digital commons enterprises and the open source development movement promoted by github and other code repositories, the commitment to open access data has been largely undertheorized. Using case studies of Digital Humanities projects that have been developed using Native American and Indigenous content, this submission suggests that Native and Indigenous content complicates the current technical application of open source development driven by digital aggregators and application programming interface development. By highlighting ethical issues around the use, reuse, and distributed architectures encouraged by common digital humanities technologies, this submission suggests that the rhetoric and practice of the open access data movement obscures both Native agency in determining the use of community materials as well as the role of technical determinism in proliferating the violence of colonial archives on Native communities. Questions this submission engages with include: How do we deal with born-digital research data in Native American and Indigenous contexts? How do we as scholars responsibly engage in digital research in Native communities? How do organizations and institutions navigate the cultural, legal, and ethical contexts of the communities whose objects they hold? How can free and open source software solutions be leveraged to build community engagement? Finally, what might be recommended for tribal communities who desire to launch their own digital projects but may have concerns about resources, access, infrastructure, and preservation?

Bibliography

Anderson, J. (2005) "Indigenous knowledge, intellectual property, libraries and archives: Crises of access,control and future utility." *Australian Academic & Research Libraries* 36, no. 2 : 83-94

Brown, D., and Nicholas, G. (2012) "Protecting indigenous cultural property in the age of digital democracy: Institutional and communal responses to Canadian First Nations and Māori heritage concerns." *Journal of Material Culture* 17, no. 3: 307-324.

Christen, K. (2011)"Opening archives: respectful repatriation." *The American Archivist* 74, no. 1 (2011): 185-210;

Christen, K. (2012). "Does information really want to be free? Indigenous knowledge systems and the question of openness." International Journal of Communication, Vol. 6, 24.

Colwell, C. (2015) "Curating Secrets." *Current Anthropology* 56: S000;

Colwell-Chanthaphonh, C. (2011) "Sketching knowledge: Quandaries in the mimetic reproduction of Pueblo ritual." *American Ethnologist* 38, no. 3 (2011): 451-467;

Cushman, E., (2013) "Wampum, Sequoyan, and Story: Decolonizing the Digital Archive." *College English* 76, no. 2: 124.

Fixico, D., (2013) *The American Indian Mind in a Linear World: American Indian Studies and Traditional Knowledge.* Routledge,

Joffrion, E., and Fernández, N. (2015) “Collaborations between Tribal and Nontribal Organizations: Suggested Best Practices for Sharing Expertise, Cultural Resources, and Knowledge,” *The American Archivist* 78:1, 193.

Madsen, D. L., (2012) *Native Authenticity: Transnational Perspectives on Native American Literary Studies*. SUNY Press.

Meyers, R. (2016) "Who Stole Native American Studies II: The Need for an AIS Redux in an Age of Redskin Debate and Debacle." *Wicazo Sa Review* 31, no. 1 : 132-144

Powell, T., and Aiken, L., (2010) "Encoding Culture: Building a Digital Archive Based on Traditional Ojibwe Teachings." *The American Literature Scholar in the Digital Age*, ed. by Amy E. Earhartand Andrew Jewell (Ann Arbor: University of Michigan Press): 250-74;

Runde, A. (2010) "The Return of Wampum Belts: Ethical Issues and the Repatriation of Native American Archchival Materials." *Journal of Information Ethics* 19, no. 1 : 33.

Senier, S. (2014) "Digitizing Indigenous History: Trends and Challenges." Journal of Victorian Culture 19, no. 3: 396-402

Warrior, R., (2014) "2010 NAISA presidential address: practicing Native American and Indigenous studies." *Journal of the Native Americanand Indigenous Studies Association* 1, no. 1 : 3-24.

GutenTag: A User-Friendly, Open-Access, Open-Source System for Reproducible Large-Scale Computational Literary Analysis

Adam Hammond
adam.hammond@utoronto.ca
University of Toronto, Canada

Julian Brooke
julian.brooke@unimelb.edu.au
University of Melbourne, Australia

Introduction

GutenTag is a cutting-edge resource that allows literary researchers of all levels of technical expertise to perform large-scale computational literary analysis. It allows users to build large, clean, highly customized worksets and then either analyse them in-system or export them as plain text or richly-encoded TEI. It has been built from the ground up by literary scholars for literary scholars: rather than relying on off-the-shelf tools poorly suited to the domain of literature, we have developed many of the components ourselves based on the specific demands of literary research. GutenTag is fully open-source, its analyses are based on entirely open corpora, and researchers can save and distribute all the parameters of their analyses, allowing for unprecedented reproducibility of research in a field plagued by siloed corpora. GutenTag is easy to use, permitting casual non-programmers to perform complex computational literary analysis via an online interface, while offering additional offline customization options to more advanced users. Although GutenTag was initially designed to facilitate our own research in polyvocality and dialogism, we show here that it can be leveraged to intervene in pressing debates unrelated to our specific research, such as the discussion surrounding Matthew Jockers's analysis of gender in *Macroanalysis*.

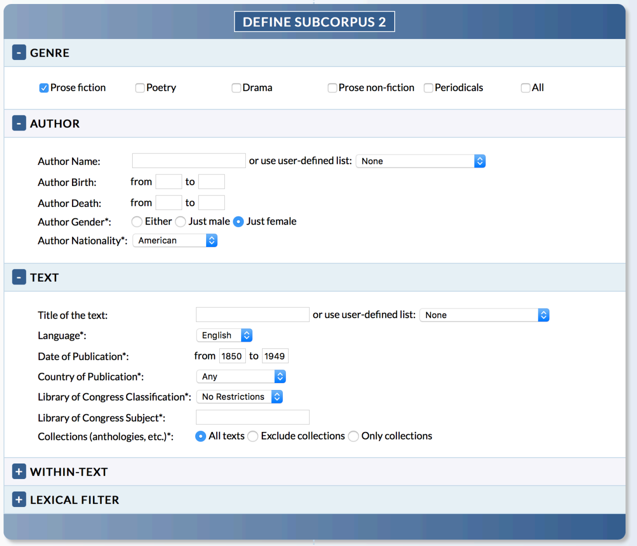
Overview of GutenTag

The system has grown considerably since our initial proposal, presented to an audience of computer scientists (Brooke et al., 2015). Below, we review the main features of the software with particular emphasis on recent improvements.

Interface: GutenTag is primarily accessed through an HTML GUI, accessible via the web or as a downloadable tool (both can be accessed from http://www.projectgutentag.org). In offline mode, the configuration files can be saved and loaded, and additional lexicons and other lists used for analysis can be specified by the user. A Python API is also included.

Corpora: The original version supported only the 2010 image of Project Gutenberg USA, but we have expanded support to all texts from Project Gutenberg USA as well as Project Gutenberg Canada and Australia, which include many additional texts published after 1922 and still under copyright in the USA.

Metadata: Document collections of interest can be defined using a variety of metadata tags. These include metadata provided by Project Gutenberg (title, author, author birth, author death, and, for some texts, Library of Congress classification and subjects). We have added genre (fiction, non-fiction, poetry, drama), determined using a sophisticated machine classifier, as well as author and text information (author gender, author nationality, publication date, publication country, single work or collection, etc.) derived from (mostly) unstructured resources including Wikipedia and the texts themselves.



DEFINE SUBCORPUS 2

- GENRE

☑ Prose fiction ☐ Poetry ☐ Drama ☐ Prose non-fiction ☐ Periodicals ☐ All

- AUTHOR

Author Name: ____ or use user-defined list: None
Author Birth: from ____ to ____
Author Death: from ____ to ____
Author Gender*: ○ Either ○ Just male ● Just female
Author Nationality*: American

- TEXT

Title of the text: ____ or use user-defined list: None
Language*: English
Date of Publication*: from 1850 to 1949
Country of Publication*: Any
Library of Congress Classification*: No Restrictions
Library of Congress Subject*: ____
Collections (anthologies, etc.)*: ● All texts ○ Exclude collections ○ Only collections

+ WITHIN-TEXT

+ LEXICAL FILTER

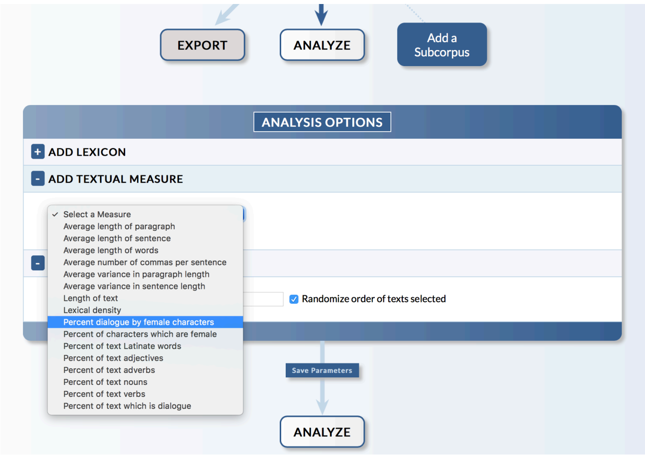
Figure 1: The GutenTag interface, showing the creation of a workset based on advanced metadata (Genre, Author Sex, Author Nationality, Date of Publication)

Text cleaning and tokenization: Sophisticated regex-based heuristics are applied to remove meta-text elements related to Project Gutenberg before, after, and sometimes within the text boundaries. Literature-specific tokenization is provided, preserving important information needed for downstream analysis.

Structural Tagging: This module identifies the main structural elements of the texts. First, heuristics are used to identify the likely boundaries between front matter, body, and back matter. Identification of structure within the main text is driven primarily by the identification of headers, and fully supports recursive structures including entire embedded texts which can have their own front and back matter separate from that of the anthology. Structural tagging is sensitive to genre: in the context of fiction, we identify parts, chapters, and speech; for poetry, we identify poems, cantos, stanzas, and lines; for drama, we identify acts, scenes, speakers, speech, and stage directions.

Lexical tagging: GutenTag includes lemmatization and POS tagging. There are several built-in lexicons which capture semantic and stylistic distinctions, and users can define their own lexicons, including multiword lexicons. Most recently, and most relevant to our case study below, we have added our own state-of-the-art literature-specific named entity recognition system (LitNER) which bootstraps from context-based clustering of common named entities to distinguish previously unseen people and locations from other named entities (Brooke et al. 2016b). For fiction, we group individual person names into collections of characters, and then assign speech events to these characters in the vicinity, using efficient, rule-based logic inspired by work in He et al. (2013). We identify the indicated sex of these characters primarily using large lists of names and titles; when a name does not appear on our list, we fall back to matching common sex-indicative character n-grams automatically derived from those lists (e.g. names ending with "a" tend to be female).

TEI output: When corpus output is required, we use XML-based TEI format as the default output format when structure (rather than simply tokens) is requested.


EXPORT
ANALYZE
Add a Subcorpus
ANALYSIS OPTIONS
ADD LEXICON
ADD TEXTUAL MEASURE
✓ Select a Measure
Average length of paragraph
Average length of sentence
Average length of words
Average number of commas per sentence
Average variance in paragraph length
Average variance in sentence length
Length of text
Lexical density
Percent dialogue by female characters
Percent of characters which are female
Percent of text Latinate words
Percent of text adjectives
Percent of text adverbs
Percent of text nouns
Percent of text verbs
Percent of text which is dialogue
Save Parameters
ANALYZE

☑ Randomize order of texts selected

Figure 2: The GutenTag interface, showing in-system options for analysis via textual measure

Analysis: In addition to building corpora for exporting, GutenTag users can directly compare the distribution of relevant lexical tags or other textual metrics across multiple corpora as defined in the metadata filtering phase. The latest version includes a selection of standard textual metrics (e.g. average sentence length), part-of-speech based metrics such as lexical density, and metrics that rely on structural/lexical tagging, such as the amount of dialogue and the amount of dialogue that has been assigned to female characters. Advanced users can easily define their own textual metrics using Python; these then become available through the main interface. We also welcome requests for metrics from the DH community.

Research Applications

GutenTag was initially developed to facilitate our own research in literary dialogism (Hammond et al. 2016, Brooke et al. 2016a). GutenTag allows us to perform three crucial steps in our research process: first, to build customized corpora (a set of novels published from 1880-1950, for which it yields 4,088 results); second, to identify passages of character speech in each novel and assign a unique character to each passage of speech; and third, to calculate a measure of dialogism for each text using an algorithm based on our six-style approach (Brooke et al. 2016a). Further, GutenTag allows us to save our workflow in a parameter file so that it can be reproduced by other researchers.

GutenTag is designed as a general system, however — not merely as a vehicle for our specialized research. We thus present an example of how it can be employed (by a non-programmer) to investigate a prominent debate in Digital Literary Studies, Matthew L. Jockers's discussion of gender and authorship in Macroanalysis. Jockers argues that female authorship can be predicted reliably through topic modelling, based on the presence of themes that "correspond rather closely to our expectations and our stereotypes" such as "Affection and Happiness," "Female Fashion," and "Infants" (Jockers 2013). A reader might respond to Jockers's analysis by querying his assumptions about literary authorship; specifically, his failure to distinguish between authors and characters. Suppose that female characters were just as likely to discuss "Female Fashion" in novels written by men as those written by women, but that female authors tended to include more female character speech in their novels, as Muzny et al. (2016) suggest. If this were so, Jockers's findings would not confirm stereotypes about female authorship, but simply reveal the tendency of female authors to include more female voices in their texts than men.

GutenTag is uniquely suited to investigating such a question. Its advanced metadata and sophisticated lexical tagging allow it to easily and rapidly analyze the question of female character speech in a large corpus of English-language novels.

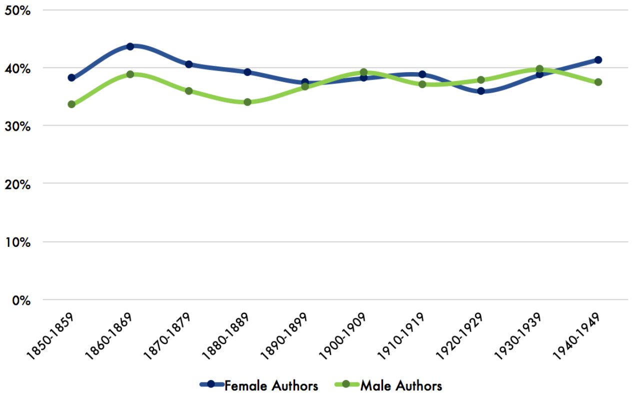


Figure 3: Mean proportion of text which Is dialogue in prose fiction, female vs. male authors, 1850-1949.

Sample sizes as follow, in number of texts. 1850-1859: 53 female, 97 male. 1860-1869: 86 female, 128 male. 1870-1879: 110 female, 137 male. 1880-1889: 122 female, 262 male. 1890-1899: 221 female, 583 male. 1900-1909: 299 female, 975 male. 1910-1919: 354 female, 960 male. 1920-1929: 148 female, 656 male. 1930-1939: 77 female, 413 male. 1940-1949: 52 female, 135 male.

Figure 3 shows that female authors in the twentieth century included approximately the same amount of dialogue as a proportion of total text length as male authors, but that in the latter half nineteenth century, they included approximately 5% more than men. Since Jockers focuses on the nineteenth century, this finding alone might impact his conclusions.

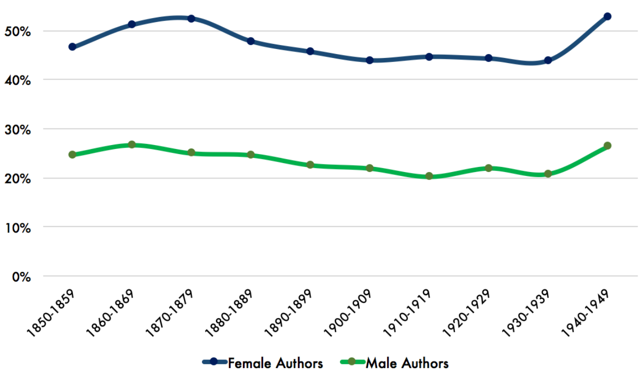


Figure 4: Mean proportion of dialogue allotted to female characters in prose fiction, female vs. male authors, 1850-1949

Sample sizes as follow, in number of texts. 1850-1859: 53 female, 97 male. 1860-1869: 88 female, 128 male. 1870-1879: 110 female, 137 male. 1880-1889: 122 female, 261 male. 1890-1899: 220 female, 583 male. 1900-1909: 300 female, 795 male. 1910-1919: 354 female, 960 male. 1920-1929: 148 female, 655 male. 1930-1939: 77 female, 413 male. 1940-1949: 54 female, 135 male.

As Figure 4 shows, GutenTag supports Muzny et al.'s contention that female novelists incorporate far more (approximately twice as much) female dialogue compared with male novelists. The finding that the proportion of female dialogue decreased from the late nineteenth to the mid-twentieth century, in both female and male authors, is one that bears further investigation — particularly in relation to the emergence in that period of popular genres, such as children's literature, Westerns, and romance novels.

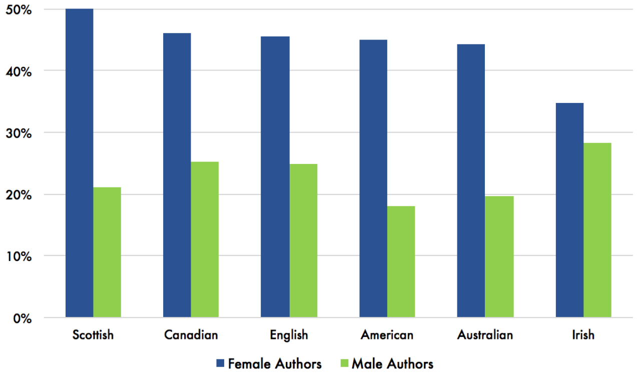


Figure 5: Mean proportion of dialogue allotted to female characters in prose fiction, female vs. male authors, by nationality, 1850-1949

Sample sizes as follow, in number of texts. Scottish: 31 female, 80 male. Canadian: 49 female, 78 male. English: 339 female, 1308 male. American: 572 female, 1545 male. Australian: 38 female, 104 male. Irish: 21 female, 92 male.

In Figure 5, we employ GutenTag's ability to filter results by author nationality. The marked discrepancy between proportion of female dialogue in male authors from England and the United States again suggests the need for an further investigation of genre; for instance, whether the American preference for male-centred genres like the Western might explain the result. Looking at GutenTag's fine-grained outputs, we observe that the texts with the lowest proportion of female dialogue are those directed at a young male audience (especially adventure fiction for boys) while those with the highest proportion consist largely of fiction for young women (L. M. Montgomery's Anne of Green Gables devotes over 90% of its dialogue to female characters). These findings might prompt our hypothetical researcher to engage in a smaller-scale study of the representation of gender in children's literature. Because all texts in GutenTag are accessible to users, it easily accommodates such movements from large-scale analysis to close reading.

Conclusion

GutenTag allows researchers of all levels of technical expertise to perform advanced large-scale literary analysis, as well as to independently test the hypotheses and conclusions of prominent research in the field. Our case study further shows how the integrated, end-to-end GutenTag system allows users to raise new research questions in the course of their analyses (such as the correlation between the emergence of children's fiction and the proportion of female dialogue) and then, since all its corpora are accessible, to shift scales and explore these questions through close reading.

Bibliography

Brooke, J., Hammond, A., and Hirst, G. (2016a). Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction. *Digital Scholarship in the Humanities*, 2(2): 1–17

Brooke, J., Hammond, A., and Baldwin, T. (2016b). Bootstrapped Text-level Named Entity Recognition for Literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL '16)

Brooke, J., Hammond, A., and Hirst, G. (2015). GutenTag: an NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. *Workshop on Computational Linguistics for Literature*. Denver: NAACL, pp. 1–6.

Hammond, A., Brooke, J. (2016). Project Dialogism: Toward a Computational History of Vocal Diversity in English-Language Literature. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 543-544.

He, H., Barbosa, D. and Kondrak, G. (2013). Identification of Speakers in Novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL '13).

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History.* Urbana-Champaign: University of Illinois Press.

Muzny, G., Algee-Hewitt, M., Jurafsky, D. (2016). The Dialogic Turn and the Performance of Gender: the English Canon 1782-2011. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 296-299.

Don’t Get Fooled by Word Embeddings—Better Watch their Neighborhood

Johannes Hellrich
johannes.hellrich@uni-jena.de
Friedrich Schiller University Jena, Germany

Udo Hahn
udo.hahn@uni-jena.de
Friedrich Schiller University Jena, Germany

Word embeddings, such as those created by the word2vec family of algorithms (Mikolov et al., 2013), are the current state of the art for modeling lexical semantics in Computational Linguistics. They are also getting more and more popular in the Digital Humanities, especially for diachronic language research (see below). Yet the most common methods for creating word embeddings are ill-suited for deriving qualitative conclusions since they typically involve random processes that severely limit the reliability of results—repeated experiments differ in which words are deemed most similar with each other (Hellrich and Hahn, 2016a,b). We provide a short overview of different embedding methods and demonstrate how this lack of reliability might affect the outcome of experiments. We also recommend a more recent embedding method, SVD$_{\text{PPMI}}$ (Levy et al., 2015), which seems immune to these reliability problems and, thus, much better suited (not only) for the Digital Humanities (Hamilton et al., 2016).

Word embeddings are a form of computational distributional semantics for determining a word’s meaning "from the company it keeps" (Firth, 1957, p. 11), i.e., the words it co-occurs with. The word2vec algorithms have their origin in heavily trimmed artificial neural networks. Their skip-gram negative sampling (SGNS) variant is widely used because of its high performance and robustness (Mikolov et al., 2013; Levy et al., 2015). Two other word embedding methods were inspired by word2vec: GloVe (Pennington et al., 2014) tries to avoid the opaqueness stemming from word2vec’s neural network heritage through an explicit word co-occurrence table, while the more recent SVD$_{\text{PPMI}}$ (Levy et al., 2015) is built upon the classical pointwise mutual information co-occurrence metric (Church and Hanks, 1990) enhanced with pre-processing steps and hyper-parameters from the two aforementioned algorithms.

There are two sources of randomness affecting the training of SGNS and GloVe embeddings: First, the random initialization of all word embedding vectors before any examples are processed. Second, the order in which these examples are processed. Both can be replaced by deterministic alternatives, yet this would simply replace a random distortion with a fixed one, thus providing faux reliability only useful for testing purposes. In contrast, SVD$_{\text{PPMI}}$ is conceptually not affected by such reliability problems, as neither random initialization takes place nor is a relevant processing order established.

Word embeddings can be compared with each other to measure the similarity of words (typically by cosine)—an ability by which they are often assessed (see e.g., Baroni et al. (2014) for more details on their evaluation). In the Digital Humanities, they have already been used to directly track diachronic changes in word meaning by comparing representations of the same word at different points in time (Kim et al., 2014; Kulkarni et al., 2015; Hellrich and Hahn, 2016c; Hamilton et al., 2016). They can also be used to track clusters of similar words over time and, thus, model the evolution of topics (Kenter et al., 2015) or compare neighborhoods in embedding spaces for preselected words (Jo, 2016). Besides temporal variations, word embeddings are also suited for analyzing geographic ones, e.g., the distinction between US American and British English variants (Kulkarni et al., 2016). In most of these approaches, the local neighborhood of selected words in the resulting embedding spaces, i.e., words deemed to be most similar with a word in question, are used to approximate their meaning at a given point in time or in a specific domain. Yet the aforementioned randomness leads to a lack of replicability, since repeated experiments using the same data set and algorithms result in different neighborhoods and might thus mislead researchers.

To investigate this problem, we trained three models each with three embedding methods, i.e., GloVe and SVD$_{\text{PPMI}}$, on the same data set and measured how they differ in their outcomes on word neighborhoods. Our data set consists of 645 German texts from the 19th century that are part of the *Deutsches Textarchiv Kernkorpus* (DTA) [German text archive core corpus] (Geyken, 2013; Jurish, 2013). The DTA contains manually transcribed texts selected for their representativeness and cultural importance; we use the orthographically normalized and lemmatized version, with casefolding. We evaluate the word embedding methods by calculating the percentage of neighbors for the most frequent nouns in the DTA on which all three models of each method agree. Overall, SVD$_{\text{PPMI}}$ provides perfect reliability, while the other two embedding methods lack reliability, SGNS dramatically so, which is consistent with our prior studies on word2vec (Hellrich and Hahn, 2016a,b).
Figure 1 shows the reliability for each model evaluated against the 1000 most frequent nouns in the DTA when their first ten closest neighbors (from one up to ten) are compared. Larger neighborhood size had a small positive effect on the reliability of SGNS and GloVe, yet is clearly unable to mitigate the inherent unreliability of these methods. A small inverse effect can be observed when the number of

the most frequent nouns is modified while keeping a constant neighborhood size of five, as displayed in Figure 2. Finally, Table 1 provides differing neighborhoods for *Herz* [heart] as a qualitative example. In this case, though not necessarily in general, SGNS models featured a more anatomical view (e.g., *bluten* [to bleed]), whereas GloVe models uncovered metaphorical meaning (e.g., *gemüt* [mind]) and SVD$_{\text{PPMI}}$ came out with a mix thereof. Using SGNS or GloVe models to assess a word's meaning can be strongly misleading, as evidenced by e.g., three SGNS models representing three different runs under the same experimental set-up. They lead to completely different semantic characterizations of *Herz* [heart], since two provide negatively connotated words (e.g., *schmerzen* [pain]) as closest neighbors, whereas the third provides a more positive impression (e.g., *herzen* [to caress]).

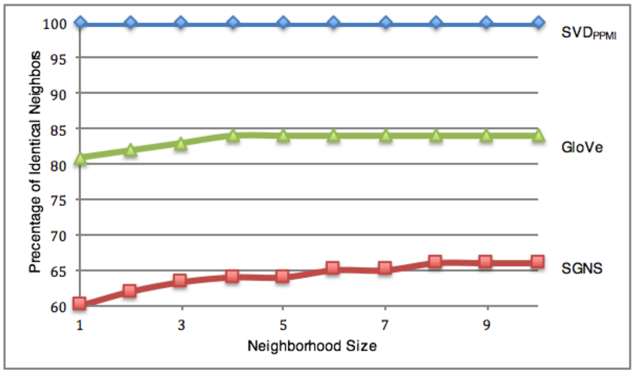


Figure 1: Reliability of different word embeddings as percentage of identical neighbors among the one to ten closest neighbor(s) to the 1000 most frequent nouns.

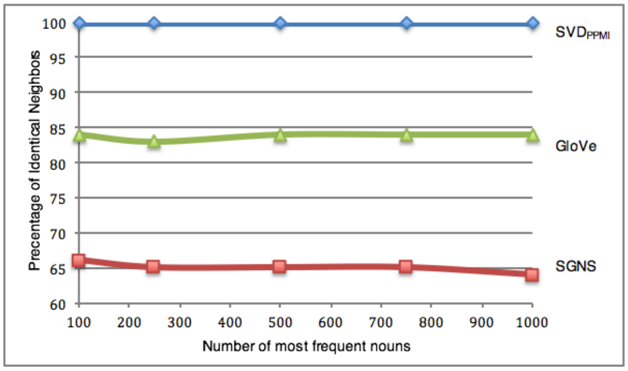
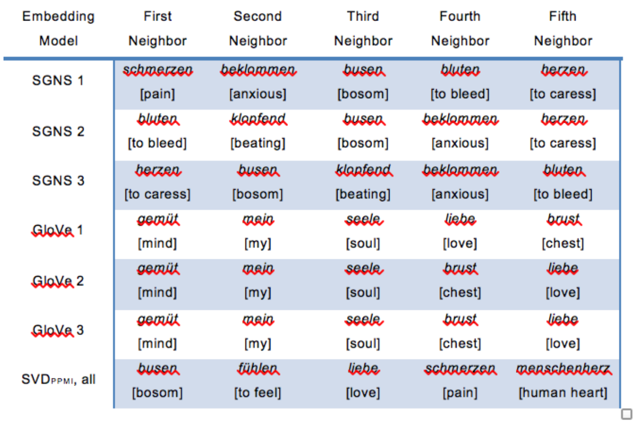


Figure 2: Reliability of different word embeddings as percentage of identical neighbors among the five closest ones for the 100 to 1000 most frequent nouns.



Embedding Model	First Neighbor	Second Neighbor	Third Neighbor	Fourth Neighbor	Fifth Neighbor
SGNS 1	*schmerzen* [pain]	*beklommen* [anxious]	*busen* [bosom]	*bluten* [to bleed]	*herzen* [to caress]
SGNS 2	*bluten* [to bleed]	*klopfend* [beating]	*busen* [bosom]	*beklommen* [anxious]	*herzen* [to caress]
SGNS 3	*herzen* [to caress]	*busen* [bosom]	*klopfend* [beating]	*beklommen* [anxious]	*bluten* [to bleed]
GloVe 1	*gemüt* [mind]	*mein* [my]	*seele* [soul]	*liebe* [love]	*brust* [chest]
GloVe 2	*gemüt* [mind]	*mein* [my]	*seele* [soul]	*brust* [chest]	*liebe* [love]
GloVe 3	*gemüt* [mind]	*mein* [my]	*seele* [soul]	*brust* [chest]	*liebe* [love]
SVD$_{\text{PPMI}}$, all	*busen* [bosom]	*fühlen* [to feel]	*liebe* [love]	*schmerzen* [pain]	*menschenherz* [human heart]

Table 1: Neighborhoods for Herz [heart] as provided by different word embedding models.

The lack of reliability we observed is definitely problematic, as often, especially for illustrations, rather small neighborhoods are used to gauge a word's meaning. Our experimental data lead us to caution when SGNS or GloVe word neighborhoods are used for uncovering lexical semantics. We recommend SVD$_{\text{PPMI}}$ instead, as its results are of similar quality yet guaranteed to be reliable (Levy et al., 2015; Hamilton et al., 2016). Consequently, we adapted our ongoing research activities on tracking language change to these insights and replaced the results of earlier work with SGNS (Hellrich and Hahn, 2016c) by data based on SVD$_{\text{PPMI}}$ (Hellrich and Hahn, 2017).

Acknowledgements

This research was conducted within the Graduate School "The Romantic Model" supported by grant GRK 2041/1 from the Deutsche Forschungsgemeinschaft (DFG).

Bibliography

Baroni, M., Dinu, G. and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pp. 238–47.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, pp. 1–32.

Geyken, A. (2013). Wege zu einem historischen Referenzkorpus des Deutschen: das Pro- jekt Deutsches Textarchiv. *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pp. 221– 34.

Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pp. 1489–501.

Hellrich, J. and Hahn, U. (2016a). An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. *Proceedings of the 10th SIGHUM Work-*

shop on Language Technology for Cultural Heritage, Social Sciences, and Humanities @ ACL 2016, pp. 111–7.

Hellrich, J. and Hahn, U. (2016b). Bad company—Neighborhoods in neural embedding spaces considered harmful. *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 2785–96.

Hellrich, J. and Hahn, U. (2016c). Measuring the dynamics of lexico-semantic change since the German Romantic period. *Digital Humanities 2016*, pp. 545–7.

Hellrich, J. and Hahn, U. (2017). Exploring Diachronic Lexical Semantics with JeSemE. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jo, E.S. (2016). Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP. *Digital Humanities 2016*, pp. 582–5.

Jurish, B. (2013). Canonicalizing the Deutsches Textarchiv. In Hafemann, I. (ed.), *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. pp. 235–44.

Kenter, T., Wevers, M., Huijnen, P. and de Rijke, M. (2015). Ad hoc monitoring of vocabulary shifts over time. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1191–200.

Kim, Y., Chiu, Y., Hanaki, K., Hegde, D. and Petrov, S. (2014). Temporal analysis of language through neural language models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61–5.

Kulkarni, V., Al-Rfou, R., Perozzi, B. and Skiena, S. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*. pp. 625–35.

Kulkarni, V., Perozzi, B. and Skiena, S. (2016). Freshman or fresher? Quantifying the geographic variation of language in online social media. *Proceedings of the 10th International AAAI Conference on Web and Social Media*, pp. 615–8.

Levy, O., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3: 211–25.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pp. 3111–9.

Pennington, J., Socher, R. and Manning, C.D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–43.

Prosodic Clustering via Cosine Similarity of Sound Sequence Inventories

Christopher Hench
chench@berkeley.edu
UC Berkeley, United States of America

Much of the discourse on music and rhythm before the time of the Minnesänger ('love singers', composers of German love poetry and songs around 1200) considers Latin chant. Alberic of Monte Cassino, active one century earlier than the Minnesänger, was the first to address rhythmic poetry in a music theory context in his treatise *De rithmis*, but he crucially does not mention rhyme, preferring to distinguish differing treatments of syllable length (Davis, 1966; Fassler, 1987). Not until the 12th and 13th centuries do treatises emerge elevating rhyme as integral to rhythm (Mari, 1899). A clear trend emerges in the period directly preceding MHG's *Blütezeit* (intense period of literary production)—theorists deemphasize syllable length in favor of a greater emphasis on syllable count and rhyme. After the Minnesänger, the *Meistersänger* ('master singers') of the German early modern period believed they were continuing the famed tradition of the medieval *Minnesänger* by focusing their art on syllable count and placement (März, 2000). For the Meistersänger, this focus was less a reflection of the increasing emphasis on music in poetic texts beginning in the 14th century, and more a method to reproduce the work of the medieval poets as closely as possible.

Unfortunately, we know very little about rhythm and sound in vernacular German poetry between these Latin treatises and the songs of the Meistersänger. Yet as both surrounding periods point to the syllable as fundamental to the composition of rhythmic poetry, we suggest the syllable as a rich source of formal information, both aesthetic and stylistic, which can serve as a formal alternative to common lexical methods in quantitative analyses, and which can help disambiguate the tension between form and content when subsequently compared to lexical methods. Although the broader project encompasses several analyses of sound and rhythm, this paper focuses on one application in particular—contrasting and visualizing the different uses of prosodic sound in the medieval German corpus (written in a stage of the language referred to as Middle High German (MHG)) through cosine similarity measurements of prosodic feature sequence *tfidf* (term frequency inverse document frequency) matrices.

Fortunately, although no unambiguous accounts of MHG rhythm survive, the manuscripts do provide phonologic evidence. We know the language of MHG had vowels of varying sonority, was structured in syllables, and was composed of words, regardless of the dialect or orthography (Paul et al., 2007). While we know that long vowels, and thus varying heavy syllables, did exist, we cannot be sure how the authors intended them in verse. Metrical systems have also been theorized, but orthographic variation in long vowels and theoretical disputes allow only limited, though productive, investigations (Estes and Hench, 2016).

To consider rhythm and sound broadly across the entire MHG corpus we turn to a relatively simple NLP technique in calculating the cosine similarity between *tfidf* inventories of texts. Yet our 'terms' are not words, but rather syllable features. For guidance, we turn to biology and new methods for clustering DNA sequences (Volkovich et al. 2005; Tomović et al. 2006; Maetschke et al. 2016). A technique employed in this scholarship takes n-gram samples of DNA strands. Similarly, we propose taking n-gram samples of *prosodic* features, primarily syllable features (closed 'C', open 'O', and word boundaries '-'), and constructing something resembling a *tfidf matrix.* Because our syllabification methodology is a combination of the sonority sequencing principle (Jesperson, 1904) and onset maximization and legal initials (Venneman, 1995), our syllabification method does not bias a dialect or orthography, but is accurate for most variants of MHG (Estes and Hench, 2016). Our features are coded as below:

Ein ritter sô gelêret was, (*Der Arme Heinrich*, l. 1-2)
C-CC-O-OOC-C-X

daz er an den buochen las,
("There was a knight so learned, that he read in the books")
C-C-C-C-OC-C-1

Where 'C' is a closed syllable (ends in a consonant), and 'O' is an open syllable (ends in a vowel). Hyphens for word boundaries account for the stress-initial tendency of MHG. Numbers at the end of a line mark end-rhyme; the number is how many lines back the rhyme was seen and an 'X' stands for the beginning of a rhyme pair (it was not seen in the past lines). All sequences for a text are then joined:

C-CC-O-OOC-C-X-C-C-C-C-OC-C-1 [...]

N-grams with an *n* of 10 are taken from this long string, going between lines for coherency, resulting in a tfidf feature matrix with feature strings of length 10. Most choices of *n* yield similar results; a lower *n* simply results in a higher degree of similarity between every text. With every increase of *n*, this similarity inevitably decreases. An *n* of 10 allows for sequences of around three to four words to be compared across texts.

A match between the MHG epics *Parzival* and *Tristan* illustrates these features:

Ist zwîvel herzen nâchgebûr
("If the heart lives with doubt",) (*Parzival* l. 1)

C-OC-CC-COC-Xvon sinen schulden ungemach
([which had] suffered due to him) (*Tristan* l. 769)
C-OC-CC-COC-X

This match implies that the number of words and syllables per word are the same, the syllable quality patterning is the same, and importantly, the rhyme is the leading rhyme (in a pair). Because the sequence matches for 13 features including word boundaries, this will create three additional matches in the tfidf inventories when the 10-gram samples are taken. While these two lines also happen to share the same scansion in the Heusler tradition, we cannot assume that every match also bears the same scansion, as more than phonology dictates metrical value (Heusler, 1956). A visualization of these cosine similarity relationships between 595 verse texts from the *Mittelhochdeutsche Begriffsdatenbank* is available online (MHDBDB, 2016).'

One may object: what if this method does not abstract *enough*? What if the syllable sequences texts share are exact lexical matches? In order to determine to what degree this prosodic sequencing approach is lexically driven, we undertake two separate measures: 1) correlation and rank correlation between the suggested formal method and a traditional lexical method, and 2) on the basis of two sample texts, we remove every possible lexical match in the inventory of sound DNA matches via a Levenshtein distance threshold, and recalculate the cosine similarities.

To account for genre intertexualities or formulas affecting this measure we take a prototypical Arthurian romance *Iwein* and its 10 most similar texts. When considering texts between an oral and written culture it is important to recognize formulaic language as influential on genre, a field pioneered by Adam Parry and Milman Parry for Homeric verse (Parry and Parry, 1971); in the Germanic tradition Franz Bäuml (Bäuml, 1972; 1976). For each n-gram sequence match in *Iwein* to each of the 10 most similar texts, we evaluate the corresponding lexical strings of the matching prosodic sequences with the Levenshtein ratio (the Levenshtein ratio is defined by the Levenshtein distance (edit distance) divided by the alignment length), if the ratio is > .85, the prosodic sequence is removed from *Iwein*, e.g., 'wîp unde man â' ≈ 'wîp unde man ze' has a Levenshtein ratio > .85, so all sequences of '-C-CO-C-XO' are removed from *Iwein*. Removing sequences of close lexical matches in *Iwein* removes 40.65% of the prosodic sequence feature strings, yet correlation of text cosine similarities before and after removal remains high at .991, and top 10 and top 20 overlap are 60% and 70% respectively, implying that text similarities are not primarily lexically driven, though lexical similarities still account for many of the mutual sequences.

To investigate this further and disambiguate the relationship between form and content, the correlation (Pearson's and Spearman's) is calculated between the similarity ranks of the prosodic sequencing method and ranks of a traditional lexical method using lemmata unigrams (*r* .624 (rank .640)), bigrams (*r* .799 (rank .801)), and trigrams (*r* .834 (rank .839)). While these coefficients are high, the main concern are the nearest neighbors, a top 20 overlap of the two methods reveals a slightly different picture: unigrams (21.8%), bigrams (32.6%), and trigrams (36.2%). These results suggest that while form and content in MHG together contribute to what one may call ‘genre’, a large share of this grouping may be similarities in form derived from the prosodic sequencing features.

Which texts exhibit the most similar and most different rank similarities between the two methods? The top 20 overlap argues that the best-matched texts in form and content are Heinrich von Veldeke’s *Eneide*, Konrad von Würzburg’s *Herzmaere*, and Gottfried von Straßburg’s *Tristan*. Interestingly, most texts in the top 10 for being best-matched are those most studied by scholars historically, and are broadly considered founders of the genre. In contrast, the most mismatched texts in form and content are Konrad’s *Der Ritterspiegel*, *Die Klage der Kunst*, and the anonymous *Lohengrin* (none of the top 20 most similar texts measured by form are the same as the top 20 texts measured by content). Severe mismatch is often an intentional aesthetic strategy, made famous by Wolfram in *Willehalm* and *Titurel*, producing what Christoph März calls a “Verfremdungseffekt”, or defamiliarization effect, per Shklovsky (März, 1999: 327).

Bibliography

Aue, H. von (2004). Der Arme Heinrich. *Gregorius ; Der Arme Heinrich ; Iwein*. 1. Aufl. (Bibliothek Des Mittelalters Bd. 6). Frankfurt am Main: Deutscher Klassiker Verlag.

Bäuml, F. H. (1986). The Oral Tradition and Middle High German Literature. *Oral Tradition*, **1**: 398–445.

Bäuml, F. H. and Bruno, A. M. (1972). Weiteres zur mündlichen Überlieferung des Nibelungenliedes. *Deutsche Vierteljahrsschrift Für Literaturwissenschaft Und Geistesgeschichte*, **46**: 479–93.

Davis, H. H. (1966). The “De rithmis” of Alberic of Monte Cassino: A Critical Edition. *Mediaeval Studies*, **28**: 198–227.

Estes, A. and Hench, C. (2016). Supervised Machine Learning for Hybrid Meter. *Proceedings of the Fifth Workshop on Computational Linguistics for Literature, NAACL-HLT 2016*: 1.

Fassler, M. E. (1987). Accent, Meter, and Rhythm in Medieval Treatises ‘De rithmis’. *The Journal of Musicology*, **5**(2): 164–90.

Gottfried (1843). *Tristan und Isolt.* (Dichtungen Des Deutschen Mittelalters Bd. 2). Leipzig: Göschen.

Heusler, A. (1956). *Deutsche Versgeschichte: Mit Einschluss des Altenglischen und Altnordischen Stabreimverses*. 2 unveränderte Aufl. Vol. 2. (Grundriss Der Germanischen Philologie 8). Berlin: W. De Gruyter.

Jespersen, O. (1904). *Lehrbuch der Phonetik*. Leipzig, Teubner.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**: 707.

Maetschke, S. R., Kassahn, K. S., Dunn, J. A., Han, S.-P., Curley, E. Z., Stacey, K. J. and Ragan, M. A. (2010). A visual framework for sequence analysis using n-grams and spectral rearrangement. *Bioinformatics*, **26**(6): 737–44.

Mari, G. (1899). *I Trattati Medievali Di Ritmica Latina*. . Vol. 11. U. Hoepli.

März, C. (1999). Metrik, eine Wissenschaft zwischen Zählen und Schwärmen?. In Müller, J.-D. and Wenzel, H. (eds), *Mittelalter: Neue Wege Durch Einen Alten Kontinent*. Stuttgart: Hirzel.

März, C. (2000). Der Silben Zall, der Chunsten Grunt. Die gezälte Silbe in Sangspruch und Meistergang. *Zeitschrift Für Deutsche Philologie*, **119**(2000): 73–84.

Mittelhochdeutsche Begriffsdatenbank (MHDBDB) (2016). Universität Salzburg. Koordination: Margarete Springeth. Technische Leitung: Nikolaus Morocutti/Daniel Schlager. 1992-2017. URL: http://www.mhdbdb.sbg.ac.at/ (2016).

Parry, M. and Milman, A. (1971). *The Making of Homeric Verse: The Collected Papers of Milman Parry*. Oxford: Clarendon Press.

Paul, H., Klein, T., Solms, H.-J. and Wegera, K.-P. (2007). *Mittelhochdeutsche Grammatik*. Tübingen: Niemeyer.

Tomović, A., Janičić, P. and Kešelj, V. (2006). N-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedicine*, **81**(2): 137–153.

Vennemann, T. (1995). Der Zusammenbruch der Quantität im Spätmittelalter und sein Einfluß auf die Metrik. *Quantitätsproblematik und Metrik*. (Amsterdamer Beiträge zur älteren Germanistik 42.1995). Amsterdam: Rodopi.

Volkovich, Z., Kirzhner, V., Bolshoy, A., Nevo, E. and Korol, A. (2005). The method of N-grams in large-scale clustering of DNA texts. *Pattern Recognition*, **38**(11): 1902–12.

Wolfram, Lachmann, K., Nellmann, E. and Kuhn, D. (1994). *Parzival*. 1 Aufl. (Bibliothek deutscher Klassiker 110). Frankfurt am Main: Deutscher Klassiker Verlag.

WorldViews: Access to International Textbooks for Digital Humanities Researchers

Steffen Hennicke
schwedes@leibniz-gei.de
Georg Eckert Institute for Int’l Textbook Research
Germany

Lena-Luise Stahn
stahn@leibniz-gei.de
Georg Eckert Institute for Int’l Textbook Research
Germany

Ernesteo William De Luca
deluca@leibniz-gei.de
Georg Eckert Institute for Int’l Textbook Research
Germany

Kerstin Schwedes
Georg Eckert Institute for Int’l Textbook Research
Germany

Andreas Witt
witt@ids-mannheim.de
Institut für Deutsche Sprache, Germany

Abstract

This paper introduces the field of international textbook research and discusses how the WorldViews project is working towards enhanced access to textbook resources for digital humanities research.

Textbook Research

The field of textbook research is one of the more recent and more diverse areas of academic investigation. The condensed and canonical character of the information selected for inclusion in textbooks - here understood as conventional textbooks - gives them central significance in academic, political and educational respects. Textbooks, as carriers of the knowledge and information that one generation wishes to pass on to the next, frequently find themselves at the center of political controversy. As such, their importance as an object of investigation in historical and cultural research has gained significant traction in recent decades. However, textbook research has not yet found dedicated representation as a main subject at universities.

As a non-university institution, the Georg Eckert Institute (GEI) for International Textbook Research conducts and facilitates applied and multidisciplinary research into textbooks and educational media primarily informed by history and cultural studies. For this purpose, the GEI provides digital and social research infrastructure services such as its renowned research library and various dedicated digital information services, such as Edumeres, a virtual network with modules for specific aspects of textbook research. As such, the GEI realizes a unique position in the international field of textbook research.

The study of textbooks has not only been facilitated by growing institutional support and infrastructure but also by the proliferation of new digital methods and resources in humanities research. In the digital humanities, the investigation of research questions is supported by a range of increasingly sophisticated digital methods such as automatic image and text analysis, linguistic text annotation, or data visualization. Digital tools and services combined with the increasing amount of resources available through digital libraries (such as the German Digital Library, the Deutsches Textarchiv, and Europeana) and research infrastructures (such as CLARIN or DARIAH) provide digital support for textbook analysis.

Digital Information Services

At the GEI, the shift towards more digitally oriented research has resulted in a range of digital information services specifically tailored to textbook research. EurViews, for example, is a multilingual digital platform containing primary sources from twentieth and twenty-first century history textbooks from around the world that manifest particular concepts of Europe and Europeanness (for example, see Gehler and Vietta, 2010; Best et al, 2012, and Chakrabarty, 2000). The service also offers essays, commentaries, and educational histories written by designated experts in the field. EurViews is a useful tool for historians searching for relevant, reliable and hard-to-find research materials on topics related to textbooks. The materials may provide inspiration for research projects or be the starting point for more extensive searches for sources. EurViews also demonstrates that the printed monograph is no longer the dominant form of publication but that digital representation is gaining in importance.

Digital representations are increasingly becoming objects of investigation themselves. GEI-Digital , for example, a digital library focusing on out-of-copyright works published between the inception of textbooks in the seventeenth century and the demise of imperial Germany in 1918, holds potentially relevant textual resources for EurViews. Other important information services provided by the GEI offer factual and bibliographic data relevant for textbook research: edu.data, for example, provides information about textbook systems in individual countries, and the library catalogue contains bibliographic metadata about textbooks from around the globe.

While these services all contain resources and information that are immediately relevant to textbook research, their content is frequently stored in isolated data silos that lack appropriate interfaces or standardized data models and which prevent convenient use or exchange of data within the GEI or with external services. For example, even though GEI-Digital makes metadata about its resources available as METS/MODS encoded data via an OAI-PMH interface, EurViews lacks the interfaces that would enable it to utilize these full-text resources. Similarly, data on textbook systems from edu.data or bibliographic metadata from the library catalogue cannot be reused by existing or new research projects. External digital humanities infrastructures such as CLARIN-D are also unable to easily access texts stored in GEI-Digital or EurViews since those platforms contain plain text documents only which are not semantically annotated.

WorldViews

The WorldViews project, which is financed by the Federal Ministry of Education and Research (BMBF), addresses these challenges in order to further elevate access to textbook resources. Its aim being to enhance discoverability, reusability, and sustainability of textbook resources in cultural and historical research and in the digital humanities. The formative use case has been the digital platform for textbook sources, EurViews, a service that has proven to be in high demand– Since 2015 EurViews has had an average of 400 visitors per month, a comparatively high number for this kind of digital academic service– and which constitutes one of the cornerstones of the GEI's research infrastructure.

Two significant milestones have been achieved on the path to enhanced access. The first is the selection of the Text Encoding Initiative (TEI) standard that will be used to encode the semantics of conventional textbook resources and thus facilitate access to full-text sources. In addition, the Component MetaData Infrastructure (CMDI) framework has been chosen for the overall integration of metadata at the GEI; a framework also compatible with the CLARIN-D infrastructure.

An extensive search for publicly available and dedicated encoding schemas for textual characteristics of conventional textbooks yielded no results. Therefore, a profile specifying the most relevant metadata and textual features in textbooks was developed from scratch after consulting historians at the GEI. The only viable option for creating an encoding schema based on the profile turned out to be the TEI encoding standard. The TEI schema created for textbooks focuses on basic elements for the selective and formal description of those structural and semantic features that are immediately relevant for WorldViews. For example, headings of sections or the semantics of particular paragraphs should provide the necessary semantics to enable retrieval scenarios to contextualize search results and to formulate more precise queries targeting particular segments of the text. This ground-breaking schema is designed to provide the nucleus for more comprehensive descriptions of whole textbooks, such as those found in GEI-Digital.

The Component MetaData Infrastructure (CMDI) provides a framework that allows blueprints of distinct metadata components to be defined and reused. CMDI allows standard metadata components to act as profiles for virtually any kind of metadata. CMDI profiles have been created to describe textbooks resources in WorldViews (we have switched to the new version of CMDI, released in mid-2016) and we are currently investigating their application in more fact-based resources such as edu.data. By using CMDI as a general framework for metadata descriptions, full-text resources can immediately be indexed by CLARIN's Virtual Language Observatory and can be analyzed using its various tools and services such as Weblicht. The CMDI description of GEI resources allows for internally standardized search and retrieval operations in federated search scenarios.

The second milestone for better access to textbook resources was the implementation of a logic tier. Central components of this tier are Solr-indices for federated searches, tools for handling digitization workflows, controlled metadata annotation and full-text annotation of textbook sources and, significantly, a Fedora repository that will dynamically provide access to standardized representations of textbook resources and other data from the various digital infrastructure services at the GEI for internal as well as external consumers. After extensive evaluation of similar repository software such as DSpace, Fedora was selected due to the greater customizability and flexibility offered by its strong modularity and existing applications with CMDI metadata. Furthermore, use of Fedora is a prerequisite for becoming a CLARIN center, one of the long-term objectives of the GEI. Through the logic tier digital platforms such as EurViews are able to send queries for bibliographic metadata on newly added textbooks directly to the library catalogue or they can directly reuse textbook resources from other systems such as GEI Digital or obtain contextual data from databases such as edu.data.

Summary

The WorldViews project has four main strategies aimed at improving overall access to digital textbook resources through enhanced reusability, discoverability and sustainability. These are a logic tier based on the Fedora repository, which mediates data between internal and external services; the digitization workflow tool Goobi, which controls metadata descriptions of digitized textbook resources; full-text encoding schemas based on TEI; and the metadata standardization based on the CMDI framework. WorldViews has laid the groundwork for standardized data models at full-text and metadata level in the field of textbook research; thereby providing a firm foundation for textbook resources and related data to be made accessible for digital humanities research.

Bibliography

Gehler, M., and Vietta, S. (Eds). (2010). Europa – Europäisierung – Europäistik, Wien.

Best, H., Lengyel, G., and Verzichelli, L. (Eds) (2012). *The Europe of Elites. A Study into the Europeanness of Europe's Political and Economic Elites*, Oxford/ New York 2012

Chakrabarty,D. (2000) *Provincializing Europe. Postcolonial Thought and Historial Difference,* Princeton, New Jersey.

CLARIN. (n.d.) Component Metadata Infrastructure. https://www.clarin.eu/content/component-metadata [Accessed 24 March 2017]

CLARIN. (n.d.) Virtual Language Observatory. https://www.clarin.eu/vlo_[Accessed 24 March 2017]

Gehler, M., and Vietta, S. (Eds). (2010). Europa – Europäisierung – Europäistik, Wien.

Georg Eckert Institute for International Textbook Research. (n.d) EurViews. http://www.eurviews.eu/nc/en/en/home.html [Accessed 24 March 2017]

Georg Eckert Institute for International Textbook Research. (n.d) GEI Digital. http://gei-digital.gei.de/viewer/ [Accessed 24 March 2017]

Georg Eckert Institute for International Textbook Research. (n.d) WorldViews Project. http://worldviews.gei.de/en/ [Accessed 24 March 2017]

Word Vectors in the Eighteenth Century

Ryan James Heuser
heuser@stanford.edu
Stanford University, United States of America

Introduction

This talk explores how new vector-based approaches to computational semantics both afford new methods to digital humanities research, and raise interesting questions for eighteenth-century literary studies in particular. New semantic models known as “word embedding models” have generated excitement recently in the natural language processing and machine learning communities, due to their ability to represent and predict semantic relationships as complex as analogy. “Man” is to “woman” as “king” is to what?, one can ask of the model; “queen,” it will most likely reply. These models formulate analogical and other semantic relationships by computing mathematical vectors for words, such that, if *V(x)* denotes the vector for the word *x*, then the above analogy can be expressed as *V(woman) – V(man) + V(king) ≈ V(queen)*. Although these models have a longer history– vector space semantics dates from the ‘70s, having been first developed for the SMART information retrieval system (Salton, 1971) by Gerard Salton and his colleagues (Salton et al, 1975)” (Turney and Pantel, 2010)– new innovations in their speed and accuracy (see Note [1]) have renewed researchers’ interests—a development begun, in part, by Google, when researchers there unveiled newly efficient algorithms in 2013, packaged in software they released called word2vec. (The word2vec algorithm was originally described by Mikolov et al, 2013. It introduced the neural network to vector space semantics, providing an efficient means by which to compute word vectors. The GloVe algorithm from the Stanford NLP Group eschews the neural network approach, instead performing a novel method of dimensionality reduction on word collocation counts).

“Word vectors,” as these new methods are sometimes informally called, have already enabled published research into questions relevant to humanistic research, such as a recent landmark paper from researchers in the Stanford NLP Group into patterns of semantic change across centuries of discourse (Hamilton et al). However, unfortunately, word vectors have so far rarely appeared in research from the digital humanities community itself. Moreover, what work that does exist has so far been primarily circulated through blogs, rather than through published proceedings or articles. Ben Schmidt, for instance, has written an influential introduction to word vectors in his blog post “Vector Space Models for the Digital Humanities” (2015a), which also includes a documented R package for computing them. Also notable is his post, “Rejecting the gender binary” (Schmidt, 2015b), which uses word vectors to dissect the polysemy of words; as well as Michael Gavin’s post, “The Arithmetic of Concepts” (2015), which explores the conceptual implications of adding and subtracting word vectors.

On the whole, the current research landscape of word vectors in the digital humanities resembles the landscape of topic modeling years ago, when the original LDA algorithm (published in 2003 [Blei et al]), before appearing in landmark published DH studies such as Matt Jockers’ *Macroanalysis* (2013), was employed for humanistic research as early as 2006 by researchers working outside or tangentially to the digital humanities (Newman and Block).

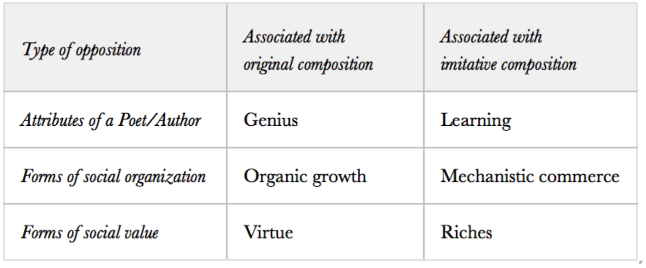
Given this scarcity of digital-humanities research on word vectors, work that seeks equally to explain, interpret, and demonstrate their potential seems particularly useful. With these goals in mind, this paper attempts first to unpack for a digital-humanities audience how word vectors work, with reference to the canonical analogy cited above: “man is to woman as king is to queen.” Second, in order to interpret word vectors’ conceptual implications for eighteenth-century literature, I move away from this canonical analogy to one central to a particularly influential argument in the period: “Learning is to Genius as Riches are to Virtue.” Lastly, I turn from this close reading of word vectors to methods of distant-reading analogies that lie implicit in eighteenth-century literature.

Explaining Word Vectors

How do word vectors work? In the interests of space, I have omitted this section of my talk from the abstract. Readers curious about the mechanics of word vectors can read more on my blog, which also links to a number of other explanatory resources (Heuser, “Methods”).

Close-reading Word Vectors

Word vectors provide a persuasive computational means for the semantic representation and analysis of analogies. They combine a mathematical elegance with an intuitive interpretability to yield what is, potentially, a method useful not only for large-scale semantic analysis, but also for smaller-scale explorations of particular analogies in literature, and their specific forms of analogical argumentation. For instance, analogy lies at the heart of Edward Young’s essay *Conjectures on Original Composition (1759),* which argued for the superior aesthetic interest of modern, “original” composition over the neoclassical imitation of the ancients. Crucially, Young makes his argument through analogy, identifying several other conceptual contrasts as analogues to his central one between original and imitative composition:

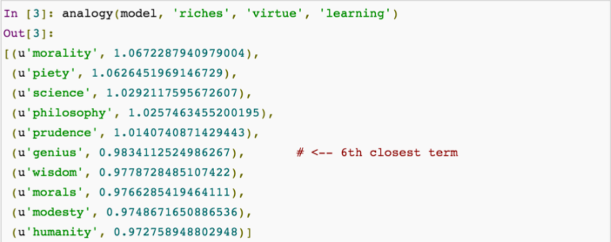


Type of opposition	*Associated with original composition*	*Associated with imitative composition*
Attributes of a Poet/Author	Genius	Learning
Forms of social organization	Organic growth	Mechanistic commerce
Forms of social value	Virtue	Riches

Table 1. Table of conceptual analogies leveraged by Edward Young to argue for original over imitative composition (Conjectures on Original Composition, 1759).

“I would compare Genius to Virtue, and Learning to Riches,” Young writes; “[a]s Riches are most wanted where there is least Virtue; so Learning where there is least Genius.” In this way, Young’s valuation of “Genius” over “Learning,” and of original over imitative composition, become *ethically justified* through their analogy with another, more obviously moral contrast between “Virtue” and “Riches.”

But what is the logic behind this analogy? Here, word vectors provide the close reader with a framework, language, and method of exploring the semantic implications at work in an analogy. In terms of vectors, we can ask, what does *V(Virtue)-V(Riches) (*Also sometimes expressed here, in a shorthand, as V(Virtue-Riches) mean, and is it in fact correlated with *V(Genius)-V(Learning)* in the broader discourse of the period? Asking this question of a word2vec model trained on the 80 million words of eighteenth-century literature in the ECCO-TCP corpus, we find that "Riches" are to "Virtue" as "Learning" is to...



```
In [3]: analogy(model, 'riches', 'virtue', 'learning')
Out[3]:
[(u'morality', 1.0672287940979004),
 (u'piety', 1.0626451969146729),
 (u'science', 1.0292117595672607),
 (u'philosophy', 1.0257463455200195),
 (u'prudence', 1.0140740871429443),
 (u'genius', 0.9834112524986267),       # <-- 6th closest term
 (u'wisdom', 0.9778728485107422),
 (u'morals', 0.9766285419464111),
 (u'modesty', 0.9748671650886536),
 (u'humanity', 0.972758948802948)]
```

Figure 1. "Riches" are to "Virtue" as "Learning" is to what?, asked of a word2vec model trained on 80 million words of eighteenth-century literature (the ECCO-TCP corpus).

"Genius" is the sixth closest word vector, or the sixth most likely solution, to this analogy. How to test the significance of this result is not immediately clear, but, out of tens of thousands of possibilities, it's certainly provocative: it raises the possibility that word vectors might provide computational assistance to close readings. Indeed, the other words in this list amplify the semantic profile of this analogy in a way that might help to clarify its underlying implications. For instance, the contrast between the *intrinsic* form of value in "Virtue" and the *extrinsic* form of value in "Riches" seems underscored for me by the contrast here between the extrinsic writerly attribute of "Learning," associated with an Oxbridge education, and the intrinsic attributes of morality, genius, and wisdom.

Ultimately, however, what does it mean to close-read word vectors? This is a question raised by Gabriel Recchia in a blogpost responding to my interpretation above as it first appeared on my blog (Recchia; Heuser, "Concepts"). Recchia's post explores other vector operations that even more reliably yield "genius," namely *V(learning)+V(virtue)* and *V(talents)+V(abilities)+V(erudition).* To me, however, these alternative "paths" to genius do not exclude one another; instead, each contributes to our understanding of the semantics of genius in the period. My goal with this interpretation is not to "prove" Young's analogy, but rather to suggest that, by "amplifying" a particular analogy through its semantic associations across a corpus, word vectors help contextualize our interpretations of particular analogies in literature. As Recchia writes, "the computational exercise has helped us focus our search."

Distant-reading Word Vectors

If, then, vectors help us explore this *micro*-analytic scale of interpretation, they also help us scale those same interpretive models up to the level of macroanalysis. For instance, inspired by the foregoing close-reading of Young's complex web of analogies (Table 1), we might continue Young's project of obsessive analogization by way of a distant reading. By defining vectors for a range of common eighteenth-century contrasts (Table 2), and then measuring the correlation between them, we can in fact construct *another* complex web of analogies—this time gleaned computationally, from a large-scale archive of the period's discourse.

Ancient(s) <> Modern(s)	Private <> Public
Beautiful <> Sublime	Romances <> Novels
Body <> Mind	Ruin <> Reputation
Comedy <> Tragedy	Simplicity <> Refinement
Folly <> Wisdom	Tradition <> Revolution
Genius <> Learning	Tyranny <> Liberty
Human <> Divine	Virtue <> Honour
Judgment <> Invention	Virtue <> Riches
Law <> Liberty	Virtue <> Vice
Marvellous <> Common	Whig <> Tory
Parliament <> King	Woman <> Man
Passion <> Reason	

Table 2: Common eighteenth-century contrasts, each expressed as a vector contrast. For instance, Virtue <> Vice denotes the vector V(Vice-Virtue). Contrasts were gleaned manually while reading Fielding's Tom Jones (1748), as well as a number of essays from the period; they are not meant to be exhaustive. This is an admittedly unsatisfactory method; I am currently exploring ways to discover conceptual contrasts computationally.

Looking at a particularly strong correlation among the contrasts in Table 2, between *V(Simplicity-Refinement)* and *V(Virtue-Vice),* we can see how their correlation emerges from the way in which both contrasts carry similar semantic associations across the same set of words (Figure 3).

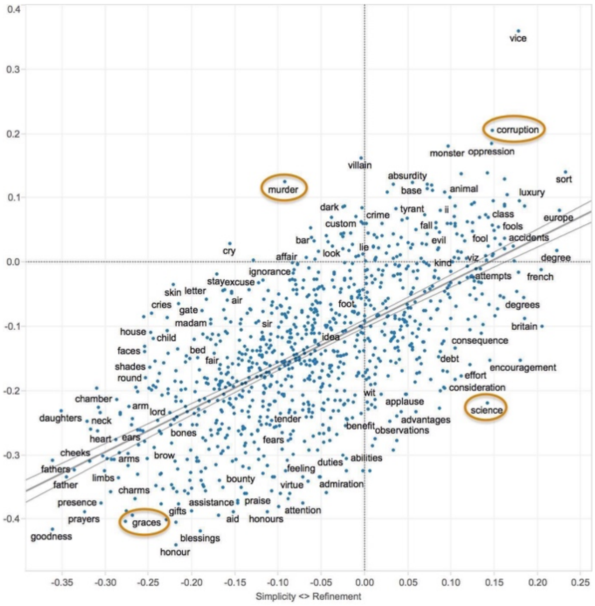


Figure 2. 1,000 most frequent nouns in the ECCO-TCP corpus. On the x-axis is their cosine similarity with the V(Simplicity-Refinement) vector: if above 0, then associated more with refinement; if below, more with simplicity. Conversely, on the y-

axis, above 0 means associated more with Vice; below 0, more with Virtue.

In other words, this graph shows that there are more words for simple virtues (e.g. "graces") than refined virtues (e.g. "science"), and more words for refined vices (e.g. "corruption") than simple vices (e.g. "murder"). This correlation between their semantic associations (R^2 = 0.41) reveals, then, an analogy emerging from the period's broader discursive practices—Simplicity is to Refinement as Virtue is to Vice—even as that analogy might appear only implicitly in particular essays, such as in Hume's "Of Simplicity and Refinement in Writing" (1742), when Hume loosely associates refinement with the moral decline of post-Augustan Rome.

This macro-analytic approach to discovering implicit discursive analogies allows us to visualize the ways in which the frequent conceptual contrasts in eighteenth-century literature are implicitly analogized in the discourse, and how those implicit analogical relationships may have helped to structure what Peter De Bolla has called the "conceptual architecture" of the period (Figure 4).

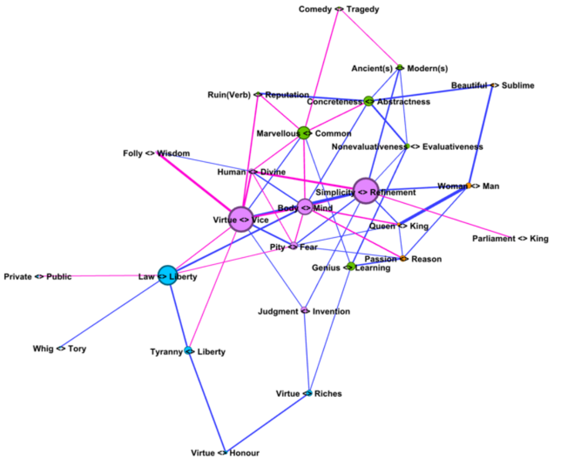


Figure 3. Semantic contrasts are connected in this network if the R^2 value of their correlation, across the 1,000 most frequent nouns (as in Figure 3), is greater than 0.1. Blue lines read in the natural order (e.g. Simplicity is to Refinement as Woman is to Man); red lines read in reverse order (e.g. Simplicity is to Refinement as the King is to Parliament). Nodes are sized by betweenness centrality, and colored by network community. Edges are sized by the R^2 value.

From this network of correlated contrasts, we can see which of them, for instance, are implicitly gendered in the period's discourse. "Woman" is to "man," for instance, as "queen" is to "king"—but also as the beautiful is to the sublime, as simplicity is to refinement, and as passion is to reason. Similarly, we can see which contrasts are moralized in the period: "virtue" is to "vice" as wisdom is to folly, as pity is to fear, as the mind is to the body. Moreover, the contrasts of virtue and vice, and simplicity and refinement, might actually play a central role in such a conceptual architecture of analogies, as seen from their centrality within the network.

Conclusion

I hope to have demonstrated some of the ways in which word vectors might be useful for the digital humanities, and particularly for eighteenth-century literary studies, both by demonstrating how they might help us to close-read specific analogical maneuvers, as well as distant-read analogies as they emerge from patterns in their usage across a literary discourse.

Notes

[1] According to statistics provided in the original paper for the Stanford NLP group's "GloVe," a competing algorithm to word2vec, a word2vec model trained on a large English-language corpus can accurately solve 65% of analogies in a test dataset, and GloVe 75% (Pennington et al, Table 2). As a rough comparison to the accuracy we would expect from human subjects, we might look to the Miller Analogy Test from Pearson—an admittedly unrelated analogy test, which is given to some graduate student applicants. In the MAT of 2002-3, to accurately solve 65% or more of its 100 analogies places a student above the 80th percentile (Pearson). Although not directly comparable, these statistics make more probable the assessment that word vectors are capable of capturing semantic relationships at a level competitive with human subjects.

Bibliography

Blei, D., Ng, A., and Jordan, M. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3.4–5 (2003): 993–1022.

Bolla, P. D. (2013) *The Architecture of Concepts: The Historical Formation of Human Rights*. Fordham UP.

Newman, D., and Block, S. (2006). "Probabilistic topic decomposition of an eighteenth-century American newspaper." *Journal of the American Society for Information Science and Technology* 57.6 (2006): 753–767.

Gavin, M. (2015) "The Arithmetic of Concepts: a response to Peter de Bolla." *Modeling Literary History*. 19 Sep 2015. Web. Accessed 1 Nov 2016.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." arXiv preprint arXiv:1605.09096. Submitted 30 May 2016. Web. Accessed 1 Nov 2016.

Heuser, R. (2016a). "Word Vectors in the Eighteenth Century, Episode 1: Concepts." *Adventures of the Virtual.* 14 Apr 2016. Web. Accessed 7 Apr 2017. <http://ryanheuser.org/word-vectors-1>

Heuser, R. (2016b). "Word Vectors in the Eighteenth Century, Episode 2: Methods." *Adventures of the Virtual.* 1 Jun 2016. Web. Accessed 7 Apr 2017. <http://ryanheuser.org/word-vectors-2>

Jockers, M. (2013) *Macroanalysis: Digital Methods and Literary History*. U of Illinois P.

Mikolov, T, et al. (2013) "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26 (2013).

Pearson. (2003) "Candidate Information Booklet." *Miller Analogies Test.* Web. Accessed 1 Nov 2016. <http://images.pear-

sonclinical.com/images/pdf/milleranalogies/matcib2002_03.pdf>

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: Global Vectors for Word Representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014), 1532–1543.

Recchia, G. (2016) "'Numberless Degrees Of Similitude': A Response To Ryan Heuser's 'Word Vectors In The Eighteenth Century, Part 1.'" *Gabriel Recchia's Blog*. 11 Jun 2016. Web. Accessed 7 Apr 2017. <http://www.twonewthings.com/gabrielrecchia/2016/06/11/numberless-degrees-of-similitude-word-vectors/>

Řehůřek, R. (n.d.) "models.word2vec – Deep learning with word2vec." *gensim*. Web. Accessed 1 Nov 2016. <https://radimrehurek.com/gensim/models/word2vec.html>

Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall.

Salton, G., Wong, A., and Yang, C. (1975). "A vector space model for automatic indexing." *Communications of the ACM* 18.11, 613–620.

Schmidt, B. (2015a). "Vector Space Models for the Digital Humanities." *Bookworm*. 25 Oct 2015. Web. Accessed 1 Nov 2016.

Schmidt, B. (2015a).. "Rejecting the gender binary: a vector-space operation." *Bookworm*. 30 Oct 2015. Web. Accessed 1 Nov 2016.

Turney, P. D. and Pantel, P. (2010)"From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37: 141-188

In Defense of Sandcastles: Research Thinking through Visualization in DH

Uta Hinrichs
uh3@st-andrews.ac.uk
University of St Andrews, United Kingdom

Stefania Forlini
sforlini@ucalgary.ca
University of Calgary, Canada

Introduction

Although recent research acknowledges the potential of visualization methods in DH, the predominant terminology used to describe visualizations (prototypes, tools) narrowly focuses on their use as a means to an end and, more importantly, as an instrument in the service of humanities research. While acknowledging the broad range of possible approaches to visualization, we introduce the metaphor of the sandcastle to highlight visualization as a research process in its own right. We argue that building **visualization sandcastles** provides a holistic approach to interdisciplinary knowledge generation that embraces visualization as (1) a dynamic interdisciplinary process where speculation and re-interpretation advance knowledge in all disciplines involved, (2) a mediator of ideas and theories within and across disciplines and (3) an aesthetic provocation to elicit critical insights, interpretation, speculation and discussions within and beyond scholarly audiences. We illustrate our argument based on our own research of an exceptional literary collection.

Visualization tools vs. sandcastles

Pivotal Scene

A steering committee meeting for a large-scale DH project. The goal of this interdisciplinary project is a combined computational and literary analysis of a literary collection, which will include the development of visualizations to enable the open-ended exploration of this collection by literary scholars. As the discussion starts to focus on the intended project outcomes, questions around the visualizations and their practical and research contributions arise. What role do visualizations play as part of DH projects? What makes them a valid contribution? One committee member brings it to the point: "Are we building **tools** or **just sandcastles**?"

This question contrasts **sandcastles**—tailored, unique, often stunning yet also transient and unstable interactive visualizations—with more pragmatic, functional and transferable visualization **tools.** This framing is a provocation: these approaches are not necessarily diametrically opposed or mutually exclusive, but, rather, exist along a rich continuum. Even within one research project, the process can shift from a more transient 'sandcastle' to a more targeted instrumental approach. And yet, the preference toward the latter is evident in recent DH discussions (Gibbs et al., 2012) and in a push by funding bodies toward research with concrete, high-impact outcomes. Notably, however, visualization 'tools or prototypes' (terms typically used interchangeably) are not usually seen as research contributions in their own right (Schreibman et al., 2010) but, at best, as facilitators of research or as a way to communicate underlying research contributions. An overly pragmatic approach to visualization, and DH tool-building more generally, however, not only risks overlooking the value of the design process but also relegating computer science and design to service-based roles. What happens when we consider, as Bruno Latour has argued, that "far from fulfilling any purpose", a new technology actually "explor[es] heterogeneous universes that nothing, up to that point, could have foreseen and behind which trail new functions" (Latour, 2002: 250)? What happens when we attend to the design process—and its many detours—as a research process in and of itself?

As a relatively young research field, information visualization (InfoVis) has seen calls to carefully and critically (re-)evaluate sometimes dated assumptions (see Kosara, 2016). Similarly, despite the increasing application of visualization in diverse DH contexts (Jänicke et al., 2015), it remains a relatively new approach and a call for generalizable visualization tools—drawing on science-based use cases—may reproduce unexamined assumptions and overlook important nuances of humanistic data and inquiry that is typically of a qualitative and interpretative nature (Drucker, 2011). As Latour argues, the ways in which we represent our arguments changes the way in which we argue (Latour, 1986). Introducing visualization into literary studies, introduces new modes of knowledge production. As such, we need to engage in open-minded and open-ended explorations of visualization as a research (rather than engineering) process, paying close attention to the ways this process changes our perspectives on data and research questions. At a time when leading practitioner-theorists suggest design as central to DH (Burdick et al., 2012), we must develop a more nuanced, critical language to discuss and further engage with the wide range of design approaches, especially from fields such as InfoVis and human computer interaction (HCI) that already combine design practice and research.

The call for a broader perspective on technology design within DH is not new, but it is increasingly urgent as the pragmatic value of visualization tools risks overshadowing the profoundly fertile design process as an intellectual and cognitive practice or a "method of thinking-through-practice" (Burdick et al., 2012). Previous work has discussed "tools" in DH as "experiments" or "embodiments of ideas" (Sinclair et al., 2011), advocated for prototypes as arguments in their own right (Galey & Ruecker, 2010), and highlighted visualization as a starting point to humanities research rather than a means to an end (Hinrichs et al., 2015; Forlini et al., 2015b; Hinrichs et al., 2016). Furthermore, critical perspectives from within the DH (Drucker, 2011) and InfoVis communities (Dörk et al., 2013; Hullman & Diakopoulos, 2011) call for further examinations of the rhetoric of visualizations. Expanding on these discussions, we reclaim the 'sandcastle' as a lens through which to critically examine current DH discussions of technology design and to promote an open-ended, speculative and process-oriented approach to visualization design based on a robust model of interdisciplinary collaboration that advances knowledge in all research fields involved. Our argument is grounded in critical theory, design research, HCI and InfoVis, as well as in our own experience of combining research in literary studies and visualization to explore an untapped collection—the Gibson Anthologies of Speculative Fiction (Forlini et al., 2015b; Hinrichs et al., 2015).

Building sandcastles at the intersection of literary studies and InfoVis

Our project—the Stuff of Science Fiction—explores a vast untapped collection of 10,000+ science fiction stories single-handedly compiled into 888 hand-crafted anthologies by the avid science fiction fan, artist and collector Bob Gibson (see Fig.1). This unusual collection raises a number of questions regarding the evolution of science fiction in the context of popular periodicals and the role of fan practices in sustaining and promoting this popular genre. Working with a subcollection of 1,500+ stories, we developed interactive visualizations that came together as the Speculative W@nderverse (see Fig. 2) to help us explore and analyze these stories through their metadata.

The W@nderverse can be considered a tool, or at least a prototype, and we have discussed it as such in our own humanities (Forlini et al., 2015a; Forlini et al., 2015b) and InfoVis publications (Hinrichs et al., 2016). In many ways it is a means to certain valuable ends: (1) it makes the Gibson anthologies explorable from different (visual) perspectives by multiple scholarly and public audiences, (2) it has generated insights about the collection, and (3) it showcases InfoVis design considerations specific to visualizing untapped literary collections (Hinrichs et al., 2016).

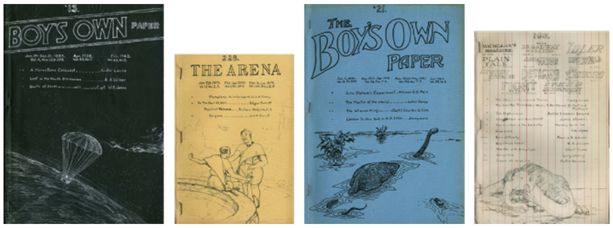


Figure 1. The Gibson Anthologies of Speculative Fiction

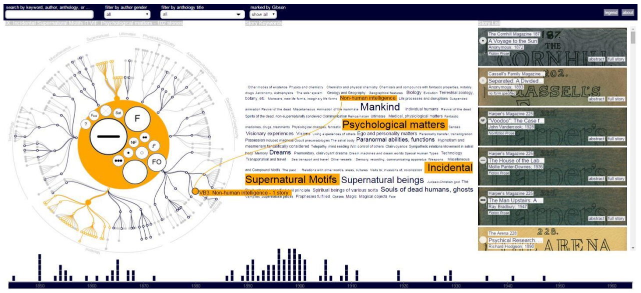


Figure 2. The Speculative W@nderverse visualization

However, if we reflect on our process with our initial research questions on one end and the visualization as a reflection of our research outcomes on the other, it becomes clear that the W@nderverse is not just a tool, at least not in the narrowly instrumental sense. It only appears to be a means to certain ends **in retrospect** when we overlook our many detours in order to narrate (for the sake of dissemination) a direct line from our questions to our contributions. However, our grant proposal and the copious notes through which we documented our research process (see Neustaedter & Sengers, 2012) remind us of our initial intentions and reveal the transformative nature of our collaborative "prototyping" process, which profoundly altered our research questions and intentions as well as our perspectives on the collection and our respective disciplines—literary studies and InfoVis. The W@nderverse is therefore both the mediator and manifestation of our exploratory and interdisciplinary research process. Our many design detours (necessitated by ongoing archival discoveries and visualization experiments, see Fig. 3), show what is now largely invisible yet fundamental to the W@nderverse: our **research thinking through visualization**, an approach that has its parallel in HCI with "research through design" (Zimmerman, 2007).

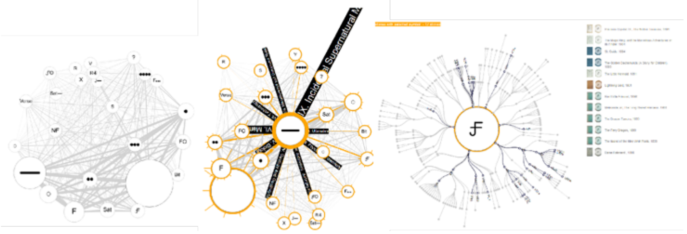


Figure 3: Early visual speculations leading up to the W@nderverse

In order to investigate humanities questions from truly novel perspectives and to engage in profoundly interdisciplinary collaborations that combine humanities and visualization research (not engineering!) approaches, we advocate for research thinking through the creation of visualization sandcastles as:

- Aesthetic and in-flux **manifestations of visualization as a speculative, provocative process** that generates insights about: 1-the underlying collection; 2-(visualization) design considerations; 3-needs of the intended audience(s); 4-and new research questions which, in turn, drive the development of new (and different) sandcastles and grounded insights valuable to all involved disciplines,
- **Dynamic mediators** that by provoking and guiding discussions can bridge boundaries between disciplines (e.g., literary studies & InfoVis) and between academic and fan endeavors, and
- **Aesthetic provocations** that can promote critical discussions of best practices for studying and making accessible cultural collections among scholarly and public audiences.

Bibliography

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., and Schnapp, J. (2012). *Digital_Humanities*. MIT Press.

Dörk, M., Feng, P., Collins, C., and Carpendale, S. (2013). "Critical InfoVis: exploring the politics of visualization." *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '13). ACM

Drucker, J. (2011). "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly (DHQ)*, 5(1).

Dunne, A., Raby. F. (2001). *Design Noir: The Secret Life of Electronic Objects*. Basel: Birkhäuser.

Galey A. and Ruecker, S. (2010). "How a Prototype Argues." *Literary and Linguistic Computing*, 25(4):405–424.

Gibbs, F. and Owens, T. (2012). "Building Better Digital Humanities Tools: Toward Broader Audiences and User- Centered Designs." *Digital Humanities Quarterly*, 6(2).

Forlini, S., Hinrichs, U., and Moynihan, B. (2015a). "Data Visualization and the Gibson Anthologies." *Presentation at MLA*, Vancouver, BC.

Forlini, S., Hinrichs, U., and Moynihan, B. (2015b). "The Stuff of Science Fiction: An Experiment in Literary History." *Digital Humanities Quarterly (DHQ); DHSI Colloquium 2014 Special Issue*, 10(1).

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., Coates, C.M. (2015). "Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration." *Digital Scholarship in the Humanities (DSH); DH2014 Special Issue*, 30(1): i50-i75.

Hinrichs, U., Forlini, S. and Moynihan, B. (2016). "Speculative Practices: Utilizing InfoVis to Explore Untapped Literary Collections." *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization, Oct. 2015)*, 22(1):429-438.

Hullman, J. and Diakopoulos, D. (2011). "Visualization Rhetoric: Framing Effects in Narrative Visualization." *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231-2240.

Jänicke, S., Franzini, G., Cheema, M.F. and Scheuermann, G. (2015). "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." *In Proc. of the Eurographics Conference on Visualization (EuroVis – State of the Art Report)*.

Kosara, R. (2016). "An Empire Built on Sand: Reexamining What We Think We Know About Visualization." *In Proceedings of Beyond time and errors: novel evaLuation methods for Information Visualization (BELIV).*

Latour, B. (2002). "Morality and Technology: The End of the Means." Trans. Couze Venn. *Theories, Culture & Society*, 19.5/6, 247-260.

Latour, B. (1986). "Visualization and Cognition: Thinking with Eyes and Hands." *Knowledge and Society: Studies in the Sociology of Culture Past and Present*, 6, pp. 1 – 40.

Neustaedter, C. and Sengers, P. (2012). "Autobiographical Design in HCI Research: Designing and Learning through Use-It-Yourself." *In Proc. of the ACM conference on Designing Interactive Systems (DIS)*, pp. 514–523.

Ruecker, S., Radzikowska, M., Sinclair, S. (2011). *Visual Interface Design for Digital Cultural Heritage - A Guide to Rich-Prospect Browsing.* Ashgate.

Schreibman, J. and Hanlon A. (2010). "Determining Value for Digital Humanities Tools: Report on a Survey of Tool Developers." *Digital Humanities Quarterly*, 4(2).

Zimmerman, J., Forlizzi, J., and Evenson, S. (2007). "Research through design as a method for interaction design research in HCI." *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, 493-502.

Beyond Validation: using programmed diagnostics to learn about, monitor, and successfully complete your DH project

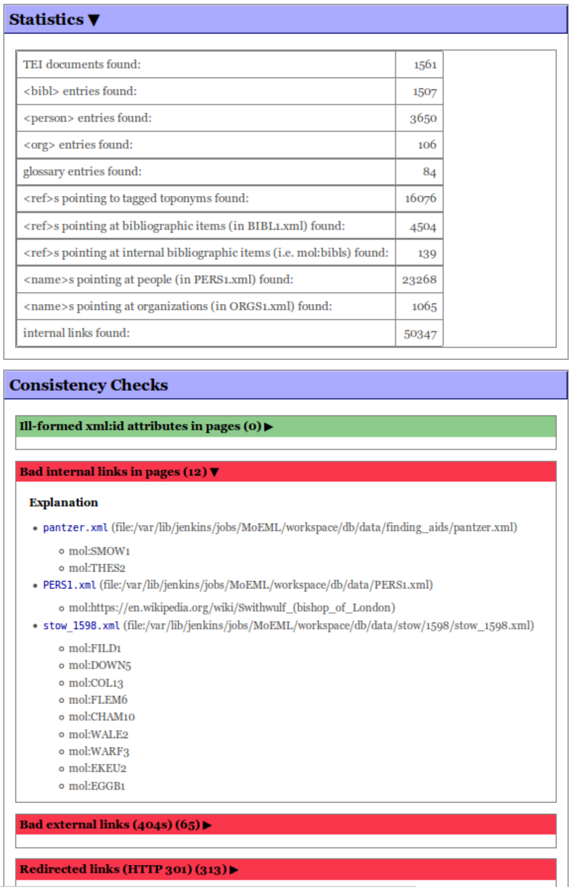
Martin Holmes
mholmes@uvic.ca
University of Victoria, Canada

Joey Takeda
joey.takeda@gmail.com
University of British Columbia, Canada

Schema-based validation of XML documents has long been a fundamental tool for quality control in digital edition projects, and the emergence of richer schema languages and adjuncts such as Schematron has greatly improved the constraints and controls available to XML authors and encoders (Jacinto et al. 2002). However, schema-based validation typically takes place at the document level, whereas "most programs that use XML require information that is not encoded in the XML instance or in the schema that governs it" (Vorthmann & Robie 2001). The modern digital edition project typically consists of multiple documents with large numbers of pointers between them: links between named entities and personographies, placeographies and bibliographies; pointers to external documents and fragments, images and other media; and similar complex interrelationships within the collection, and to external resources and authorities. These relationships need to be tested, checked, and validated too, but it is impractical to do this using document-level schemas. As Durand et al. (2009) point out, "such testing requirements are in fact closer to conventional system or software testing requirements than to document testing in a narrow sense." Most large- and medium-scale projects develop their own methods, programmed and/or impromptu, for addressing these problems, and these have been quite well-described and documented for enterprise-level and corporate contexts (for instance, see the papers presented presented at the International Symposium on Quality Assurance and Quality Control in XML, particularly Waldt 2012), but little has been published on project-level diagnostic testing for XML-based digital edition collections (see Rahtz, 2007)

In our work as part of Endings, an umbrella project that comprises four diverse digital edition projects from different fields, we have been developing a structured approach to implementing methods for checking and enforcing project correctness, consistency, and coherence, which we will describe in this paper. Influenced no doubt by Star Trek, we have long referred to these processes as "diagnostics", and in our description we follow the franchise tradition detailed in Sternbach and Okuda (1991) in dividing diagnostics into levels; however, we depart from convention in ordering our levels from most granular/least comprehensive up to the most general. For each level, we provide real examples of processes run on one of our projects.

We stress that these diagnostics are built on top of a solid basis of RelaxNG and Schematron schemas. In the case of our projects, we use highly-customized versions of the TEI schema (all TEI-compliant) in addition to project-specific Schematron rules, which not only police tagging practices (e.g. enforcing the use of private URI schemes in pointing attributes, and checking the presence of appropriate custom dating attributes for pre-Gregorian dates), but also style guide rules such as prohibiting the use of straight apostrophes in document text nodes (excepting computer code samples). Our diagnostic processes normally take the form of ant scripts and XSLT transformations, and are run on a Jenkins Continuous Integration server; every time changes are committed to a project repository, the Jenkins server checks out the changes, validates all documents, and runs the entire set of diagnostics processes, providing the results in the form of a public web page such as this one:



Statistics ▼

TEI documents found:	1561
<bibl> entries found:	1507
<person> entries found:	3650
<org> entries found:	106
glossary entries found:	84
<ref>s pointing to tagged toponyms found:	16076
<ref>s pointing at bibliographic items (in BIBL1.xml) found:	4504
<ref>s pointing at internal bibliographic items (i.e. mol:bibls) found:	139
<name>s pointing at people (in PERS1.xml) found:	23268
<name>s pointing at organizations (in ORGS1.xml) found:	1065
internal links found:	50347

Consistency Checks

Ill-formed xml:id attributes in pages (0) ▶

Bad internal links in pages (12) ▼

Explanation

- pantzer.xml (file:/var/lib/jenkins/jobs/MoEML/workspace/db/data/finding_aids/pantzer.xml)
    - mol:SMOW1
    - mol:THES2
- PERS1.xml (file:/var/lib/jenkins/jobs/MoEML/workspace/db/data/PERS1.xml)
    - mol:https://en.wikipedia.org/wiki/Swithwulf_(bishop_of_London)
- stow_1598.xml (file:/var/lib/jenkins/jobs/MoEML/workspace/db/data/stow/1598/stow_1598.xml)
    - mol:FILD1
    - mol:DOWN5
    - mol:COL13
    - mol:FLEM6
    - mol:CHAM10
    - mol:WALE2
    - mol:WARF3
    - mol:EKEU2
    - mol:EGGB1

Bad external links (404s) (65) ▶

Redirected links (HTTP 301) (313) ▶

Figure 1: A diagnostics output page from the Map of Early Modern London project.

In combination with this paper, which is intended to be a useful primer and guide, we have developed a Diagnostics project hosted on GitHub. that can be used by researchers whose digital edition projects have grown to the point where ad hoc manual checking has become impractical. This tool provides generic referential integrity checking that can be applied to any set of TEI XML files.

Level 1

Level 1 diagnostics provide project-level, as opposed to document-level, consistency checking to establish the internal coherence of the project, primarily through ensuring referential integrity. We borrow the phrase "referential integrity" from the MLA's "Guiding Questions for Vetters of Scholarly Editions" (2011), which advises peer-reviewers of digital editions that link to multiple databases to see if "referential integrity [is] enforced within the database(s)." This includes checking for non-existent pointers, duplicate @xml:ids across the project, and erroneously encoded references (e.g. tagging a place name as a bibliography reference). Ensuring referential integrity is particularly complex for projects that use "abbreviated pointers" to facilitate internal linking (see TEI Consortium (2016)), since it may not be obvious to the encoder which resource is being referenced by a pointer. Thus, the first level of diagnostics checks both whether or not an object pointed to actually exists and whether or not the markup correctly represents the relationship between the element and the target resource. For instance, to check all instances of the relationship shown in Fig. 2, a number of different tests are actually done:

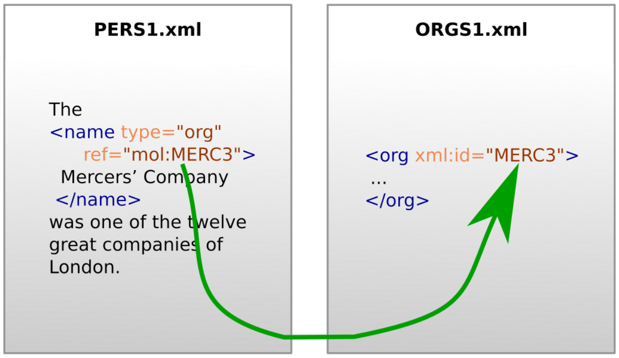


Figure 2: a simple referential integrity check.

1. Every <name type="org"> points at an @xml:id which exists in the project.
2. The element pointed at by <name type="org"> is an <org> element in the ORGS1.xml document.
3. Every <name> element which points at an <org> element in ORGS1.xml has @type="org".

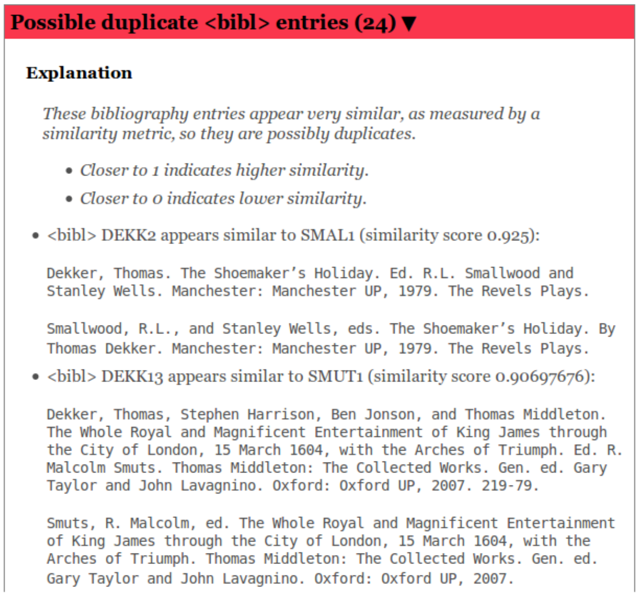
For small-scale projects, this kind of referential integrity check could be accomplished with Schematron, since a Schematron rule using XPath 2.0 can read external documents, but for a project of any significant size, this is impractical. For example, Schematron checks to confirm the rules above may add around six seconds to document validation in the Oxygen XML Editor, causing frustration for editors, while simply checking that a linked location exists would require the processing of over a thousand files in this project, since each location is a distinct file.

Level 2

While Level 1 diagnostics generally focus on coherence and consistency, Level 2 is more concerned with completeness. Level 2 diagnostics provide progress analysis, generate to-do lists, and identify situations that may indicate error, but require human judgement. These include cases in which:

- Two bibliography or personography entries appear sufficiently similar that they may be duplicates.
- Several <name> elements point to the same authority record, but the text of one of them is significantly different from the others, so it may point at the wrong target.
- A document in the project is not linked from anywhere else, and therefore cannot be "reached".

Such issues cannot be automatically rectified—they are not necessarily errors—but they must be examined. Figure 3 shows an example of the first check, which uses a similarity metric to identify potential duplicate bibliography entries.



Possible duplicate <bibl> entries (24) ▼

Explanation

These bibliography entries appear very similar, as measured by a similarity metric, so they are possibly duplicates.

- *Closer to 1 indicates higher similarity.*
- *Closer to 0 indicates lower similarity.*

- <bibl> DEKK2 appears similar to SMAL1 (similarity score 0.925):

```
Dekker, Thomas. The Shoemaker's Holiday. Ed. R.L. Smallwood and
Stanley Wells. Manchester: Manchester UP, 1979. The Revels Plays.

Smallwood, R.L., and Stanley Wells, eds. The Shoemaker's Holiday. By
Thomas Dekker. Manchester: Manchester UP, 1979. The Revels Plays.
```

- <bibl> DEKK13 appears similar to SMUT1 (similarity score 0.90697676):

```
Dekker, Thomas, Stephen Harrison, Ben Jonson, and Thomas Middleton.
The Whole Royal and Magnificent Entertainment of King James through
the City of London, 15 March 1604, with the Arches of Triumph. Ed. R.
Malcolm Smuts. Thomas Middleton: The Collected Works. Gen. ed. Gary
Taylor and John Lavagnino. Oxford: Oxford UP, 2007. 219-79.

Smuts, R. Malcolm, ed. The Whole Royal and Magnificent Entertainment
of King James through the City of London, 15 March 1604, with the
Arches of Triumph. Thomas Middleton: The Collected Works. Gen. ed.
Gary Taylor and John Lavagnino. Oxford: Oxford UP, 2007.
```

Figure 3: Results of a Level 2 diagnostic check that attempts to identify duplicate bibliography entries.

At Level 2, we also generate to-do lists for specific sub-projects, providing a set of tasks for the project team to focus on in order to reach a milestone or publish a particular document. The definition of "done" for a specific document may transcend the document itself. For instance, before we deem a particular edition of a text publishable, we may require that all authority records (people, places, publications) linked from that document are themselves complete, so the to-do list for a given document may require work in a variety of other documents in the project

Level 3

Armed with a comprehensive set of Level 1 and Level 2 diagnostics, and assuming our data is managed using a version-control repository such as Subversion or Git, we can now generate diachronic views of the project's progress. A script can check out a sequence of incarnations of the project, weekly over a period of months, for instance, and run the entire current diagnostic suite against it; we can then combine these snapshots to get a clear sense of how our work is proceeding. This also means that every time we develop a new diagnostic procedure, we can apply it to the entire history of the project to see the trajectory of project work with respect to the datapoint in question. Two examples, this time from the Nxaʔamxcín Dictionary project (an indigenous dictionary project described in detail in Czaykowska-Higgins, Holmes, and Kell (2014)) appear in Figs 4 and 5 below. Figure 4 shows the number of completed dictionary entries in orange, rising steadily over a period of 18 months, and the number of occurrences of a known problem: duplicate instances of the same gloss. These duplicates rise along with the number of entries until October 2016, when this issue was added to our diagnostics process, and the encoders were able to address it.

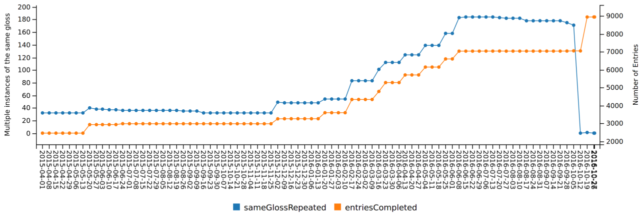


Figure 4: The number of instances of duplicate glosses, tracked against completed entries, in the Nxaʔamxcín Dictionary project.

Fig. 5 shows cases of broken cross-references, which also tend to increase along with the number of completed entries, but we can see from the graph that the issue was aggressively addressed in two separate campaigns in fall 2015 and summer 2016. New instances continue to appear, however.

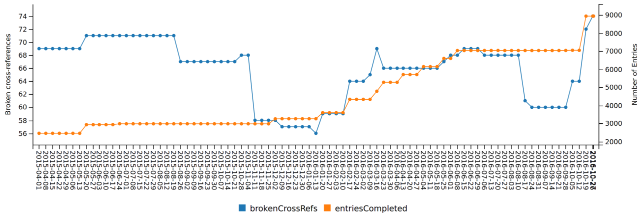
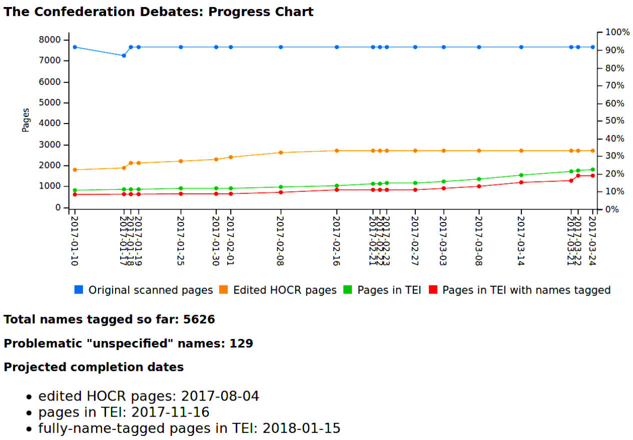


Figure 5: The number of broken cross-references, tracked against completed entries.

Fig. 6, from a different project, shows how this approach can be used to forecast completion dates for tasks in a project based on the progress rate so far.



Total names tagged so far: 5626

Problematic "unspecified" names: 129

Projected completion dates

- edited HOCR pages: 2017-08-04
- pages in TEI: 2017-11-16
- fully-name-tagged pages in TEI: 2018-01-15

Figure 6: Diachronic diagnostics used to project task completion dates.

Conclusion

As Matthew Kirschenbaum (2009) tells us, there “is no more satisfying sequence of characters” than “Done.” The overall purpose of a digital edition project is to finish and publish the edition, and this requires not only that the document-level encoding be valid, but also that the entire dataset be coherent, consistent, and complete. Programmed diagnostics enable projects to enforce coherence and consistency, manage the workflow effectively, and measure their progress towards completeness.

Bibliography

Czaykowska-Higgins, E., Holmes, M., and Kell, S. (2014). “Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project.” Language Documentation & Conservation 8: 1–37.

Modern Language Association. (2016) “Guidelines for Editors of Scholarly Editions.” Modern Language Association. Accessed September 15. https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/Guidelines-for-Editors-of-Scholarly-Editions.

Modern Language Association. (2011) “Guiding Questions for Vetters of Scholarly Editions.”. Modern Language Association. Accessed October 21. https://www.mla.org/content/download/3201/81158/cse_guidelines_2011.pdf.

Jacinto, M. H. , Librelotto, G. R., Ramalho, J. C., and Henriques, P. R. (2002). “Constraint specification languages : comparing XCSL, Schematron and XML-Schemas.” http://repositorium.sdum.uminho.pt/handle/1822/619.

Kirschenbaum, M. (2009). “Done: Finishing Projects in the Digital Humanities.” DHQ 3 (2). http://digitalhumanities.org:8081/dhq/vol/3/2/000037/000037.html.

International Symposium on Quality Assurance and Quality Control in XML (2012) “Proceedings of the International Symposium on Quality Assurance and Quality Control in XML.”

August 6, 2012. http://www.balisage.net/Proceedings/vol9/contents.html.

Rahtz, S. (2007). “Technology Overview and Discussion: Data Capture, Editing, and Schemas.” Oxford, February 13. http://tei.it.ox.ac.uk/Talks/2007-02-13-oucs/talk-editing.xml.

Sternbach, R., and Okuda, M. (1991). Star Trek, the next Generation: Technical Manual. New York: Pocket Books. http://catalog.hathitrust.org/api/volumes/oclc/24648561.html.

Vorthmann, S., and Robie, J. 2001. “Beyond Schemas: Schema Adjuncts and the Outside World.” Markup Languages: Theory & Practice 2 (3): 281–94.

Waldt, D. 2012. “Quality Assurance in the XML World: Beyond Validation.” Accessed September 15. http://www.balisage.net/Proceedings/vol9/author-pkg/Waldt01/BalisageVol9-Waldt01.html.

Transcriptional Implicature: Using a Transcript to Reason about an Exemplar

Claus Huitfeldt
Claus.Huitfeldt@uib.no
University of Bergen, Norway

C. M. Sperberg-McQueen
cmsmcq@acm.org
Black Mesa Technologies LLC
United States of America

What do transcripts tell us about their exemplars, and how? A transcript T is commonly said (inter alia) to reproduce as far as possible the letters and words of its exemplar E. The tokens of T typically reinstantiate the letter and word types of E; T may also contain commentary on E.

Transcribers vary in their practice: some record page and line breaks in E, silently omit deletions, expand abbreviations, and correct spelling; others don't. Often a statement of practice documents such details, but some things apparently go without saying. We conjecture that what goes without saying will be the most common assumptions and habits in a community of practice. We have proposed the term "transcriptional implicature" to denote the inferences licensed by a transcript although not explicitly justified in the statement of practice [SMMH2014].

Practice (and implicature) vary across communities, but we think that some silent assumptions are shared by almost every community of practice. We believe that we can identify a common core of such assumptions, and that individual practices can be characterized by listing their deviations from this core, which we call the "default rules" of transcriptional implicature.

Like all implicatures, the rules of transcriptional implicature are defeasible: they are taken to hold in the absence of evidence to the contrary. [SMMH 2014] proposes these rules:

- Reciprocity: There is a one-to-one relation between normal tokens in an exemplar E and normal tokens in its transcript T.
- Purity: Every normal token in T transcribes something in E.
- Completeness: Every normal token in E is transcribed by something in T.
- Type similarity: corresponding tokens in E and T have the same type, or non-identical but similar types.

Given these rules, the transcription practice of any project can be described by defining what counts for that project as a special (not "normal") token and what type pairs count as similar. This paper puts these general principles to an empirical test by applying them to a concrete example: the transcript, in the Northwestern-Newberry Edition, of Hermann Melville's notes on the end-papers of a volume of Shakespeare [Melville, p 967-970]

The "Symbols used" are (numbers ours):

1. [...] revision or insertion enclosed in square brackets was made later than initial inscription of leaf
2. <...> letters or words enclosed in diamond brackets were canceled by lining out
3. <...>word letter(s) or word(s) written over are enclosed in diamond brackets closed up to the following word or letter that was superimposed
4. ?word prefixed question mark indicates conjectural reading 5 xxxx undeciphered letters (number of x's approximates numbers of letters involved)
5. all words in roman are Melville's
6. all words in italics outside brackets are words Melville underlined
7. all words in italics inside brackets are editorial

If transcriptional implicature is to play the role we ascribe to it, it must be possible to recast these rules in terms of normal tokens, special tokens, and types instantiated by tokens:

- Square brackets, diamond brackets, question marks prefixed to words, and sequences of the form xxxx in T are special tokens; they transcribe no material in E.
- Italic material inside brackets in T is special.
- A word written without underlining and the same word underlined have distinct types in E; the same holds for italic and roman material in T. Words underlined in T and the same words italicized in E are type-similar.

- Later additions to E have distinctive types; that is, we can distinguish them. A sequence of words inserted in E and the same sequence enclosed in square brackets in T are type-similar.
- Overlined words, overwritten words, and words neither overwritten nor overlined instantiate different types. Overlined or overwritten word-types are instantiated in T by enclosing the words in diamond brackets.
- "Undeciphered letter sequences of length N" are distinct types for distinct N, instantiated in E by undeciphered characters and in T by sequences of the form xxxx. The individual x tokens are special: they do not transcribe tokens of E.
- A word prefixed with a question mark and the same word without it instantiate different types in T. Such a word token, and its constituent character tokens, are type-identical with their exemplars in E if the editors' conjectural reading is correct. We describe such tokens in T as "probably type-identical" to their exemplars.

Both generic and project-specific rules can be formalized using first-order logic. The generic rules include the rules of transcriptional implicature:

```
(∀t t:tokens(T))(normal-transcript-token(t) ⍰
      (∃1e e:tokens(E))(e = exemplar(t))

(∀e e:tokens(E))(normal-exemplar-token(e) ⍰
      (∃1t t:tokens(T))(t = transcript(e)))

(∀e e:tokens(E), t: tokens(T))(t = transcript(e) ⍰
      (type(e) = type(t) v type-similar(e, t))
```

Reciprocity has no formalization here; it's a property of the functions exemplar() and transcript().

Other generic rules constrain the classes of normal and special tokens:

```
(∀t t:tokens(T))(normal-transcript-token(t) |
      special-transcript-token(t))

(∀e e:tokens(E))(normal-exemplar-token(e) |
      special-exemplar-token(e))
```

The Melville-specific rules identify special tokens in T:

```
(∀t t:tokens(T))(special-transcript-token(t) ⍰
      (square-bracket(t)
      v angle-bracket(t)
      v prefixed-question-mark(t)
      v xxx-sequence(t)
      v (italic(t) & within-square-brackets(t))
      v (italic(t) & within-angle-brackets(t))
      v prefixed-question-mark(t)
      v page-furniture-token(t)))
```

The predicate page-furniture-token identifies material like running heads and page numbers in T. Project-specific type similarity rules cover conjectural readings:

```
(∀t t:tokens(T))(∀e e:tokens(E))
      (probably-identical(type(t), type(e)) ⍰
      type-similar(e, t))
```

The Melville transcription policy defines no special exemplar tokens:

```
¬(∃t t:tokens(E))(special-exemplar-token(t))
```

We hypothesize that it is rules like those above that enable readers of a transcript to draw conclusions about an exemplar they have never seen. Such reasoning can be formalized with the aid of a full formal representation of T and its mapping of tokens to types. In the case of Melville's notes, such a formalization will involve several thousand tokens in T and E and their types. The formal representation will not be given in full here, but it will include formulae like the following, in which "d", "p524", "L1", etc. are logical constants naming tokens:

```
document(d)
page(p524)
page(p525)
document_pages(d, (p524, p525)).
line_token(L1)
line_word_tokens(L1, (L1w1, L1w2, L1w3, L1w4, L1w5, L1w6, L1w7))
word_token(L1w1)
word_token(L1w2)
...
word_token_type(L1w1, "A")
word_token_type(L1w2, "seaman")
...
word_token_type(L1w7, "Tales.")
line_token(L2)
...
line_token(L40)
line_type(LT1)
line_type(LT2)
...
line_type(LT40)
page_lines(p524, (L1, L2, L3, ... L30))
page_lines(p525, (L31, L32, L33, ... L40))
```

Sequences are here represented using comma-separated lists of items, enclosed in parentheses.

As an example: We can infer from the transcript that the first line of the first page transcribed reads "A seaman figures in the Canterbury Tales." This will surprise no one who can read and understands what a transcript is. The challenge here, however, is to establish the result using nothing but the normal rules of logical inference, without hand-waving.

Let us focus on the line token L1. We draw our initial inferences from the formal representation of T. As is usual for composite types, the type of the line is determined by the types of the words which make up the line. So the type of L1 is the sequence of words (or characters) "A seaman figures in the Canterbury Tales."

```
composite_type_string(LT1, "A seaman figures in the Canterbury Tales.")
```

Since L1 is not a square bracket, diamond bracket, etc., it can be inferred that L1 is a normal not a special token. From the rule of purity it then follows that there is some token e in E which L1 transcribes. This fact, together with the rule of type-similarity, tells us that e is a token which like L1 instantiates the composite type "A seaman figures in the Canterbury Tales."

A second example: A student examining a reproduction of the original of Melville's notes might ask "Are the dots to the left of the first line significant, or just discolorations of the paper?" We take this question to mean "Do the dots instantiate a type?" Let us assume that they do. If they instantiate a type, then they are tokens in E; if they are tokens in E, then either they are normal or special. They cannot be special tokens: there are none in this transcription practice. If they are normal tokens, then there exists a token in T which instantiates the same type. But there are no such tokens in T. If there is no token in T which transcribes the dots at the upper left of the page, then they are not (in the transcribers' reading of E) normal tokens. If they are neither normal tokens nor special tokens, they are not tokens at all. So: they are merely marks, not tokens, and they instantiate no type.

We have postulated transcriptional implicature as a way of bridging the gap between the explicit statements in descriptions of transcription practice and the inferences actually licensed; the empirical test reported suggests that the concept does provide the required basis for the intended inferences. If formal representations of the information content of a transcript can be generated automatically (e.g. from a TEI-encoded transcript), then we will be slightly closer to being able to describe formally the semantics of transcription-oriented markup languages like MECS-WIT or TEI when used to transcribe pre-existing exemplars.

Bibliography

Melville, H (1998) Moby Dick, or, the Whale. Ed. Harrison Hayford, Hershel Parker, and G. Thomas Tanselle. Vol. 6 of The Writings of Herman Melville, The Northwestern-Newberry Edition 1988, rpt. 1994, 1997.

Huitfeldt, C. (1993). MECS-WIT - A Registration Standard for the Wittgenstein Archives at the University of Bergen, 1993 approx. 250 pages. Available electronically only, now at http://folk.uib.no/fafch/oldstuff/mecswit.html

Sperberg-McQueen, C. M., Marcoux, Y., and Huitfeldt, C., (2014). "Transcriptional Implicature: A Contribution to Markup Semantics." Paper at DH 2014, Lausanne. Abstract on the Web at http://dharchive.org/paper/DH2014/Paper-61.xml

TEI Consortium, eds (n.d.) TEI P5: Guidelines for Electronic Text Encoding and Interchange. [Version 3.1.0. Last updated on 15th December 2016, revision d3f5e70]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Social Semantic Annotation with Recogito 2

Leif Isaksen
l.isaksen@lancaster.ac.uk
Lancaster University, United Kingdom

Rainer Simon
rainer.simon@ait.ac.at
Austrian Institute of Technology, Austria

Elton T. E. Barker
elton.barker@open.ac.uk
The Open University, United Kingdom

Pelagios Commons (Pelagios Commons 2017) is a community of practice dedicated to supporting Linked Open Data-related activity within the humanities. Comprising a series of interconnected Special Interest Groups, it operates through the collective establishment of digital conventions for semantic data and produces software to facilitate its production and use. Such data can ultimately be used to interconnect independently maintained and heterogeneous online resources about the past. While Pelagios Commons is engaged in a variety of activities to achieve these ends, this paper focuses on the development of a second generation implementation of its most popular tool, Recogito (Simon et al., 2015). Recogito is a semantic annotation system (Andrews et al., 2012) with similarities in some regards to platforms such as Pundit (Grassi et al., 2012) and Hypothesis (Hypothesis, 2017), but with particular focus on geographic content, and accessibility to humanists without high levels of technical literacy.

Recogito's origins lie in the earlier *Pelagios 3* project cycle dedicated to the semantic annotation of early geographic documents so as to provide a 'critical mass' of content to which other historical materials could be related. The intended goal was to identify place references, and their referents, in as many pre-1500 geographic documents as possible. Such documents, or more specifically their digital surrogates, could take a wide variety of different formats, but are typically represented as texts, images and tables. While the first phases of Pelagios saw annotation carried out by a community of resource curators, much - though not all - of the annotation for Pelagios 3 was conducted within the Investigative Team. As most of the documents to be annotated had no formal structure, it was necessary to develop a tool which allowed us to rapidly produce annotation records (recording a reference, a global gazetteer entry and an annotator). The result was *Recogito*, a web-based platform which allowed its users to upload digital surrogates and produce annotations pointing back to the original digital version of an online document.

The first version of Recogito offered multiple interfaces for the discrete tasks of: identifying place references (whether in text, image or table); transcribing them where necessary; relating them to an entry in one of multiple possible gazetteers (depending on the period of the document in question). Semi-automated processes based on Natural Language Processing and Named Entity Extraction techniques allowed the software to accelerate these tasks, while ultimately requiring all annotations to be verified by a human individual. Semantic interpretation is always underdetermined by a text or symbol, and thus human intervention remains a fundamental principle of the Pelagios methodology. While document surrogates were kept behind a log-in interface to prevent the possibility of copyright infringement, the annotations themselves were made publicly available in real time under a CC0 (public domain license).

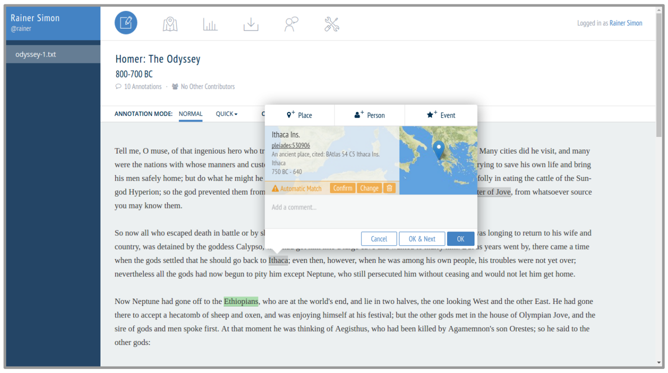
Pelagios 3 and Recogito successfully met their aims and have drawn interest from potential stakeholders working throughout the humanities. These range not only across different periods and geographic regions, but between disciplinary fields (such as archaeology, classics and history), in diverse forms of text, and for the production and alignment of gazetteers to boot. Thanks to an Open Knowledge Foundation/DM2E Open Humanities Award, Recogito was also tested with several undergraduate student classes. This not only allowed us to refine its user interface, but also to see the various ways it changed students' approaches to, and understanding of, the material they were annotating. The very process of annotation, as many digital humanists can attest, forces a level of systematic reflection upon the annotator which might otherwise be elided.

Despite - indeed because of - Recogito's success, a number of limitations came to light during the course of Pelagios 3. First and foremost was its centralised architecture which provided a single workspace for all users. While well suited to a small team in regular communication with one another, the increasing number of users and documents meant that managing document metadata, preventing errors caused by concurrent edits, and keeping users abreast of changes to documents of interest became increasingly difficult. Furthermore, administrative tasks like uploading documents could only be carried out centrally, creating unnecessary barriers to use. Above all, while document copyright was protected through a restriction on public access, it was clear that if the system user base continued to grow over time, then the risk of copyright abuse would grow with it.

In addition to this central issue were a considerable number of feature requests which were not in the project's original scope of works and which we were unable to introduce within the time available. These include support for various input and output formats (including TEI XML, KML, and GeoJSON); the ability to add non-semantically defined commentary; overlapping annotations; simple points for identifying symbols on images (rather than textboxes for toponyms); competing interpretations of place references; and the ability to make annotated documents publicly available where copyright permits. With renewed financial support from the Andrew W. Mellon Foundation, the Pelagios initiative has been able to redesign and implement Recogito from the ground up in order to address these deficiencies and introduce additional features as well.

Recogito 2 presents an entirely new interface for the semantic annotation of place references. Users can self-register and are now provided with their own workspace (with an initial storage allowance of 200MB) in which they can upload and annotate documents. Each user's cataloguing page has its own URL and is publicly visible, although any uploaded documents are not. Documents can be uploaded singly or as a batch of related files (such as images of pages within a manuscript). Documents are automatically preparsed for possible place references at the upload stage unless the user declines to do so. Currently, Recogito 2 makes use of the Stanford NLP Toolkit (Manning et al., 2014), but we intend to make it extensible so as to support alternative parsing engines, such as the Classical Language Toolkit (Johnson et al. 2014-17) which may be better suited to specific languages or use cases. Whereas Recogito 1 only supported plain text documents and image files, Recogito 2 also supports TEI XML and images held in IIIF-compliant repositories, as well as JPEG, TIFF and PNG.

Once uploaded, an 'annotation view', allows the user to confirm automatically identified place references or create new ones (Figure 1). This takes place by means of a pop-up dialog box which determines the type of annotation (currently only places references, with free-text commentary and tags, but future development will include additional support for person references and events). Assuming the reference is to a place, the system proposes a probable gazetteer candidate which can either be confirmed or corrected as appropriate. Where the same place definition has been aligned across multiple gazetteers, these will be merged into a single entity for consideration, but the user is able to select which gazetteer they wish to formally associate the reference to.



Homer: The Odyssey
800-700 BC
10 Annotations · No Other Contributors

Ithaca Ins.
pleiades:530906
An ancient place, cited: BAtlas 54 C5 Ithaca Ins.
Ithaca
750 BC - 640

Tell me, O muse, of that ingenious hero who tr
were the nations with whose manners and cust
his men safely home; but do what he might he
god Hyperion; so the god prevented them from
you may know them.

Many cities did he visit, and many
rying to save his own life and bring
folly in eating the cattle of the Sun-
ter of Jove, from whatsoever source

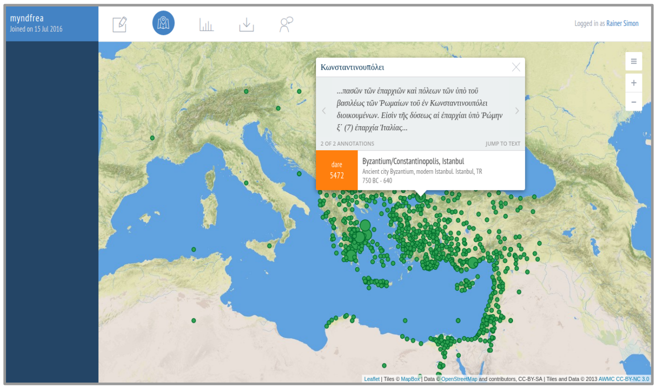
So now all who escaped death in battle or by sl
country, was detained by the goddess Calypso,
when the gods settled that he should go back to Ithaca; even then, however, when he was among his own people, his troubles were not yet over;
nevertheless all the gods had now begun to pity him except Neptune, who still persecuted him without ceasing and would not let him get home.

vas longing to return to his wife and
as years went by, there came a time

Now Neptune had gone off to the Ethiopians, who are at the world's end, and lie in two halves, the one looking West and the other East. He had gone there to accept a hecatomb of sheep and oxen, and was enjoying himself at his festival; but the other gods met in the house of Olympian Jove, and the sire of gods and men spoke first. At that moment he was thinking of Aegisthus, who had been killed by Agamemnon's son Orestes; so he said to the other gods:

Fig. 1. Recogito 2 text annotation view.

A major development in Recogito 2 is the introduction of 'social tools' for collaborative annotation. Annotators can share their documents with other registered users, allowing them to view, edit and create annotations dependent on the permission settings. Edits effectively act as discussion threads so that the full history of changes and commentary can be seen for each annotation (and where necessary, rolled back by the owner). We are also aware that for many users, producing Linked Open Data is not their primary objective. There are many other benefits to be derived from annotating place references, not least of which is the ability to map content. Recogito 2 offers a simple mapping interface that shows the distribution of place references - where coordinates can be derived from a gazetteer - against a range of possible backing maps. Symbol size reflects the number of references to the place, and selecting one provides the user with the specific references within the text itself (Figure 2).


myndfrea
Joined on 15 Jul 2016
Logged in as Rainer Simon
Κωνσταντινουπόλει
2 OF 2 ANNOTATIONS
JUMP TO TEXT
dare
5472
Leaflet | Tiles © MapBox | Data © OpenStreetMap and contributors, CC-BY-SA | Tiles and Data © 2013 AWMC CC-BY-NC 3.0

...πασῶν τῶν ἐπαρχιῶν καὶ πόλεων τῶν ὑπὸ τοῦ βασιλέως τῶν Ῥωμαίων τοῦ ἐν Κωνσταντινουπόλει διοικουμένων. Εἰσὶν τῆς δύσεως αἱ ἐπαρχίαι ὑπὸ Ῥώμην ξ΄ (7) ἐπαρχία Ἰταλίας...

Byzantium/Constantinopolis, Istanbul
Ancient city Byzantium, modern Istanbul. Istanbul, TR
750 BC - 640

Fig. 2. Recogito 2 map view.

A series of additional features replicate popular functions within Recogito 1. This includes a statistics dashboard, which provides data about both documents (such as place reference frequency, or the proportion of references lacking identification) and places (such as the toponyms by which they are referred to, and tags associated with them). It is our intention to add 'social statistics' that may allow people to see which of their documents are proving most popular or perhaps to help identify other users with similar interests. We have also extended the annotation download formats from solely CSV and RDF, to include KML, TEI and GeoJSON. This allows for their incorporation within a much wider range of software commonly used by humanists.

Already in public Beta-testing, Recogito 2 offers a next generation approach to semantic annotation of humanities documents with specific emphasis on place references. Nevertheless, we believe strongly that the value of this contribution lies not simply in its technical innovation but in facilitating community contributions to the Web of Linked Open Data.

Bibliography

Andrews, P., Zaihrayeu, I., Pane, J. (2012) "A Classification of Semantic Annotation Systems." In *Semantic Web*, 3(3): 223-248.

Grassi, M., Morbidoni, C., Nucci, M., Fonda, Ledda, G. (2012) "Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries."In *Proceedings of the 2nd International Workshop on Semantic Digital Archives* (SDA 2012).

Hypothesis. 2017. http://hypothes.is/

Johnson, K. P. et al. (2014-2017). CLTK: The Classical Language Toolkit. DOI 10.5281/zenodo.60021

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., McClosky, D. (2014) "The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Pelagios Commons. 2017. http://commons.pelagios.org

Simon, R., Barker, E., Isaksen, L., and de Soto Cañamares, P. (2015) "Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito." In *e-Perimetron.* 10(2): 49-59. ISSN.

Untangling the Social Network of Musicians

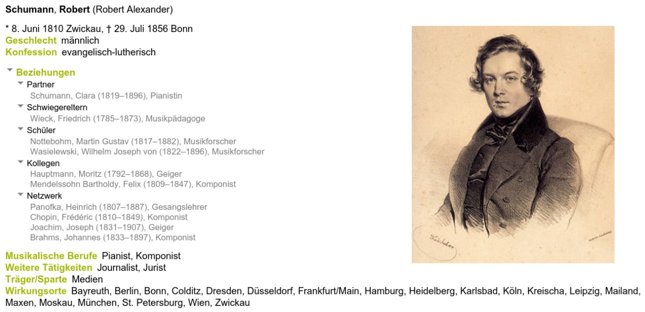
Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

Josef Focht
josef.focht@uni-leipzig.de
Leipzig University, Germany

Introduction

As opposed to the former idea of creative autonomy, in recent years, humanities research tends to investigate cultural contexts and circumstances, inspirational models, and the ways that knowledge, experience and expertise have been transferred over time. We address the question of "creative transfer" within the field of music. Due to the everlasting significance of musical works, relationships between musicians – the entry point for such an investigation – are well documented in archives, libraries and museums. In print media, usually only a single relation between two musicians is narrated. Furthermore, it is common for the biography of only one of the two musicians to report on the relationship. Larger overviews of social networks between several musicians seldom exist. Although some digital resources exist, these are often reduced to the milieus of popular musicians like Mozart and Beethoven.

Since 2005, musicologists of the project Bavarian Musicians Encyclopedia Online (Bayerisches Musiker Lexikon Online, BMLO) have systematically collected biographical data (an example is given in Figure 1) and examined relationships between musicians from print media – a tedious work that results in a unique database of great value for musicology. The BMLO contains musicians from all kinds of musical professions (e.g., composers, singers, musicologists, instrument makers, ...), most of whom had an active lifetime period living in Bavaria or a considerable influence on Bavaria. Now providing information about around 28,000 musicians, the BMLO has achieved global scope, one that is underpinned by the many musicologists worldwide who use the BMLO for their daily work.



Schumann, **Robert** (Robert Alexander)

* 8. Juni 1810 Zwickau, † 29. Juli 1856 Bonn
Geschlecht männlich
Konfession evangelisch-lutherisch

Beziehungen
- Partner
  - Schumann, Clara (1819–1896), Pianistin
- Schwiegereltern
  - Wieck, Friedrich (1785–1873), Musikpädagoge
- Schüler
  - Nottebohm, Martin Gustav (1817–1882), Musikforscher
  - Wasielewski, Wilhelm Joseph von (1822–1896), Musikforscher
- Kollegen
  - Hauptmann, Moritz (1792–1868), Geiger
  - Mendelssohn Bartholdy, Felix (1809–1847), Komponist
- Netzwerk
  - Panofka, Heinrich (1807–1887), Gesangslehrer
  - Chopin, Frédéric (1810–1849), Komponist
  - Joachim, Joseph (1831–1907), Geiger
  - Brahms, Johannes (1833–1897), Komponist

Musikalische Berufe Pianist, Komponist
Weitere Tätigkeiten Journalist, Jurist
Träger/Sparte Medien
Wirkungsorte Bayreuth, Berlin, Bonn, Colditz, Dresden, Düsseldorf, Frankfurt/Main, Hamburg, Heidelberg, Karlsbad, Köln, Kreischa, Leipzig, Mailand, Maxen, Moskau, München, St. Petersburg, Wien, Zwickau

Figure 1: Biographical information about Robert Schumann in the BMLO. Alongside information about a musician's lifetime, denomination, professions or places of activity, the database provides a number of relationships by type. Next to his partner Clara Schumann, further relations are listed to Robert Schumann's father in law, to his students, colleagues, and other acquainted musicians in his social network

In earlier works, we developed visual interfaces on the basis of the BMLO data for profiling musicians (Jänicke et al, 2016), and for the distant reading of musicians' biographies (Khulusi et al, 2016). However, the social network inherent in the BMLO has remained untouched so far. Using the BMLO, only the social network of single musicians can be observed, as is the case when using print media. In order to facilitate an extensive analysis of the entire social network concealed in the BMLO, we designed a visualization that brings together all of the relationships in the form of an interactive social network graph. In contrast to previous means of investigating the transfer of musical knowledge, we allow for the dynamic exploration of relationships among musicians over generations.

Graph Topology

Information regarding relationships to other musicians in the database is provided for 9,805 musicians of the BMLO. Only one relation exists for around 46,5% of these musicians, and just 261 musicians have ten or more relations. Adolf Wilhelm August Sandberger is the musician with most relations (97). The average number of relations for musicians is 2.6. The resultant graph structure of the social network consists of 1,420 connected components, the largest component connects 5,539 musicians, the second largest only 56 musicians – 1,385 connected components contain less than ten musicians.

Due to the above mentioned topological features of the graph, the typical, straightforward visualization using a force-directed layout approach, e.g., by using tools such as Gephi (Bastian, 2009), leads to a global overview of the social network (see Figure 2). However, local structures are hardly readable, which makes an interactive exploration nearly impossible. The objective of this work was to develop a graph design that makes the social network of musicians visually accessible for the first time, and, moreover, capable of being explored in accordance with the research questions of the collaborating musicologists. We focused on the largest connected component that causes the greatest challenges for this task.

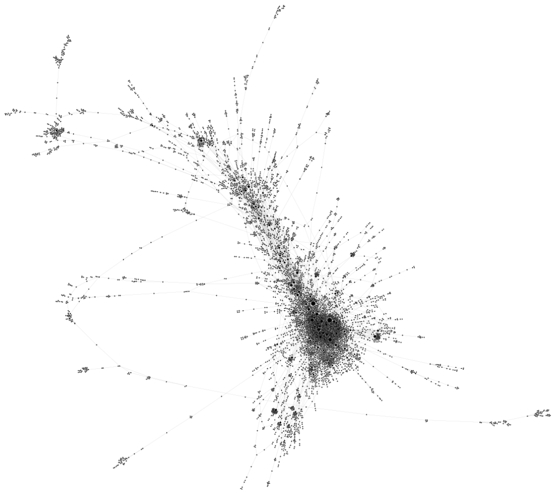


Figure 2: The largest connected component with 5,539 musicians visualized using Gephi.

Graph & Interface Design to Analyze Teacher–Student Relationships

The preliminary step when generating the social network graph is filtering according to the underlying research question. At first, a filtering can be done by relationship type(s). Second, it is possible to focus exclusively on musicians with specific professions (e.g., instrumentalists). In the following discussion, we focus on the motivating example for this work: the analysis of teacher-student relationships to investigate how musical knowledge, experience and expertise have been transferred over time. The corresponding filter keeps 3,994 musicians, the largest connected component of this sub-network – the research object of the musicologists – contains 2,769 teachers and students. The Gephi output for this graph is given in Figure 3.

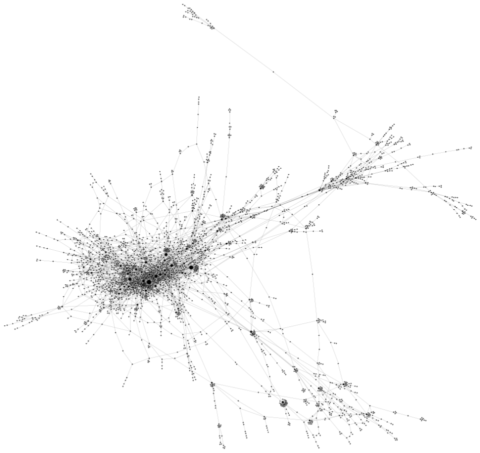


Figure 3: The largest connected component of the teacher-student network with 2,769 musicians visualized using Gephi.

Although the structures are slightly finer due to the reduced number of nodes and edges, the highly connected part in the interior of the graph remains cluttered. Here, we list our design decisions applied in order to generate a readable graph (see Figure 4) and a navigable interface.

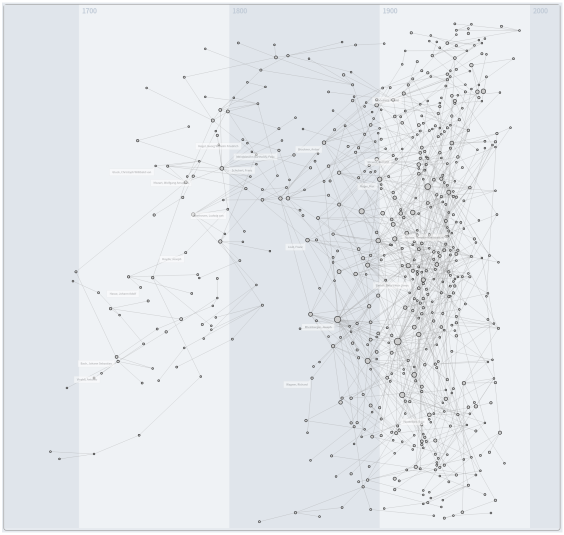


Figure 4: The largest connected component of the teacher-student network with 2,769 musicians (608 nodes) visualized with our method.

- **Temporally aligned graph:** It was particularly important for the musicologists that the graph layout includes a temporal dimension, so that relations can be chronologically analyzed from left-to-right. Therefore, we applied a force-directed graph layout and used fixed x-values that represent a time-stamp, which reflects the middle of a musician's creative lifetime (see Jänicke 2016), on a horizontal time axis. As a result, the nodes only spread vertically and the chronological order remains intact.
- **Node grouping:** Because the underlying research question investigates transfer paths of musical knowledge, we hide the nodes of musicians who never had the role of a teacher. Still, these musicians are grouped to their teachers, and can be accessed in the exploration process. This design decision reduces the number of nodes to be displayed from 2,769 to 608.
- **Node layout:** To illustrate the significance and the influence of personalities, the sizes of nodes reflect the number of students of the corresponding teachers, which makes teachers with many students salient. Per default, node labels are hidden, but for navigation purposes, a user-defined number of node labels with the corresponding musicians names can be shown on demand. Either the most popular musicians or the teachers with most students can be highlighted.
- **Interactivity:** Hovering over a node shows the corresponding musician and two lists of students (those who became teachers and those who did not) in a popup box. Clicking a node highlights all connections to a teacher's students who became themselves teachers. This way, transfer paths of musical knowledge can be assembled interactively.
- **Musical profession analysis:** For the selected (via mouse click) musicians in the graph, the evolution of musical professions can be analyzed. Therefore, all musical professions of the teachers' students are listed by decreasing frequency. For each profession, a bar chart illustrates when they have been pursued.

Analysis of Teacher–Student Relationships

This section outlines a usage scenario of the teacher-student network taking the example of Adolf Wilhelm August Sandberger who established musicology as a subject of study in Munich.

First, we compare Sandberger to one of his teachers, Joseph Rheinberger, both being the teachers with the highest numbers of students (the BMLO lists 97 students for Sandberger and 87 students for Rheinberger). Of special interest was the comparative analysis of the musical professions of their students in order to assess the similarity of both teachers' studentries. Figure 5 shows the two selected teachers in the social network, and a view at the summarized musical professions of their students is given. While composition was the major musical profession of Rheinbergers students (70x), this number drops for Sandbergers students (52x). On the other hand, the number of musicologists increase (10x → 65x). Other significant changes can be seen for the professions choirmaster (29x → 9x), organist (19x → 8x), music writer (12x → 29x) and music editor (5x → 26x). Thus, the visualization reflects a change of the musical profile of both studentries from composition to composition science– a hypothesis that could be verified with our system.

Second, we examined the change of teaching since Sandberger established musicology in Munich. Therefore, we observed the musical professions of the students of Sandberger and his successors in Munich, Rudolf von Ficker, Thrasybulos Georgios Georgiades and Theodor Göllner (see Fig. 6). While the musicologist is the most frequent taught musical profession, the composer gets less and less important. The last teacher Theodor Göllner even had no student with the composer as musical profession. Thus, the change from composition to composition science that already started with Sandberger compared to Rheinberger, steadily continued with Sandberger's successors.

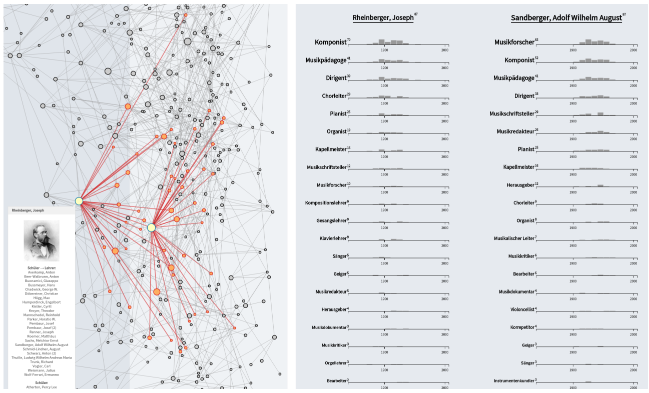


Figure 5: Comparing the students of Joseph Rheinberger and Adolf Wilhelm August Sandberger.

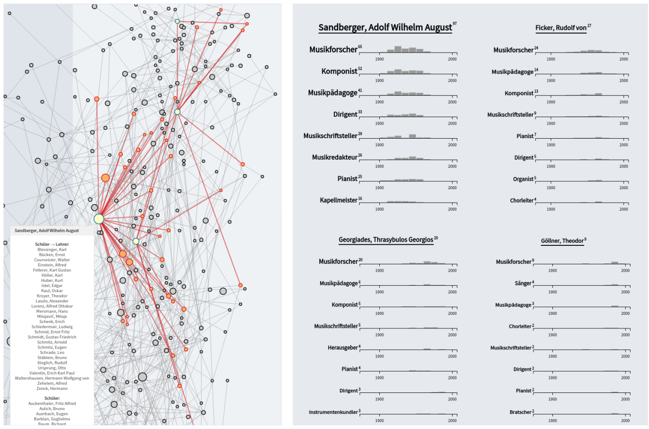


Figure 6: Temporal change of teaching in Munich.

Conclusion

Through close collaboration with computer scientists and musicologists implementing a user-centered design approach, we developed a visualization that allows for the dynamic, interactive exploration of the social network of musicians, focusing primarily on teacher-student relationships. In contrast to out-of-the-box tools like Gephi, we took the research questions of the collaborating musicologists into account when designing the graph and the user interface. Although detailed information about individual relation periods between musicians as well as the taught musical professions are not included in the underlying database, the provided interface facilitates a novel view on the social network of musicians, which allows to draw conclusions on the question of the transfer of musical knowledge.

The value of our system for users of the BMLO is not only that social networks are visualized for the first time, but also that the graph may be filtered in accordance with the way that specific research questions can be investigated. Next to teacher-student relationships, familial or labor relationships also create valuable networks to be explored. Furthermore, it is possible to analyze sub-networks concerning musical professions, and to combine relationship types with musical professions. For example, when combining teacher-student relationships with the musical profession *instrumentalist* (see Fig. 7), Wolfgang Amadeus Mozart shows up at the beginning of the instrument playing knowledge transfer.

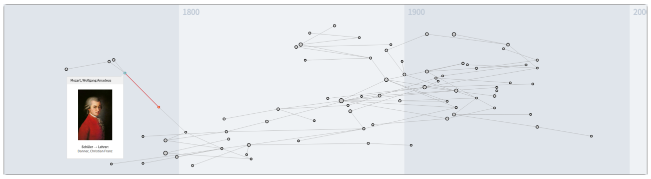


Figure 7: Teaching instrumentalists.

Bibliography

Bastian M., Heymann S., and Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Jänicke, S., Focht, J., and Scheuermann, G. (2016). Interactive Visual Profiling of Musicians. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(1):200–209, Jan 2016.

Khulusi, R., and Jänicke, S. (2016). On the Distant Reading of Musicians' Biographies. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 818-820.

Beaudouin, V., and Pehlivan, Z. (2016). The Great War on the Web: the Making of Citing and Referencing by Amateurs. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 433-436

Mapping the STC with MoEML and DEEP

Janelle Auriol Jenstad
jenstad@uvic.ca
University of Victoria, Canada

Tye Landels-Gruenewald
tyelandels@gmail.com
Queen's University, Canada

Joseph Takeda
joey.takeda@gmail.com
University of British Columbia, Canada

Introduction

MoEML's gazetteer of 6500 London place name variants invites the mapping of datasets with a geographical component. As a textual editing project with an interest in print culture, we have long hoped to mobilize our GIS tools and gazetteer data in the service of mapping the English book trade. Our ultimate goal is to publish a layer showing the printing and/or retailing locations of the approximately 25,000 books printed in London between 1475 and 1640.

Imprint lines in early modern books include highly granular location data, which has meant that book history has traditionally had an implicit geospatial dimension. A typical imprint line tells us that copies of a folio are "Printed by Elizabeth Purslovv, and are to be sold by Nicholas Bourne, at his shop at the south entrance of the Royall Exchange, 1633." Using the information in such imprint lines, Kathleen Pantzer reorganized the items in the Short Title Catalogue under location headings (Pantzer numbers) in Vol. 3 of the catalogue. Her work facilitates questions about the proximity of one printer or bookseller to another, and thereby questions about affiliations, collaborations, and specialization among a key group of early modern cultural agents.

However, considerable processing of Pantzer's printed lists is required to visualize or map STC items. Thus far, digital databases like *Early English Books Online* (EEBO) and the *English Short Title Catalogue* (ESTC) have captured the imprint line without parsing it into discrete data points, thereby leaving Pantzer's formidable interpretive work behind as we move into the era of digital historical bibliography. *The Database of Early English Playbooks* (DEEP) has included and corrected Pantzer numbers, but only for the printed plays, of course. MoEML has attempted to replicate Pantzer's work via datamining the ESTC. After several unsuccessful NER experiments on ESTC data, we are now mobilizing the curatorial work of DEEP and planning to extend their work beyond playbooks. In this paper, we take imprint lines and geospatial information about the book trade as a case study in mining carefully curated data. We explain the long history of this project as it extends back to Pantzer's own work creating the strict vocabulary for the print locations of early modern texts. We then discuss how MoEML has been able to put the STC data onto the Agas map, giving a better sense of the spatial relationship of printed early modern texts. In doing so, our argument centers on the necessity for authority names and strict vocabularies. Invoking Mike Poston's suggestion that we cannot predict the uses of our data, we use our own work on various print and digital databases to show how we can control and scaffold the mining processes to establish links between several pairs of projects in order to mine and ingest data from databases that do not share a common data field with the initial project in the sequence. We conclude with a list of considerations and principles for maximizing future interoperability between literary datasets.

Methodology

Although not strictly based in MySQL technology, our methodology borrows from the work of digital humanists like Harvey Quamen and Jon Bath who use MySQL to design relational databases. Indeed, in order to establish valuable connections across diverse datasets, we must first identify what data points these datasets have in common (either directly or indirectly). For example, suppose that Dataset 1 contains raw data for categories A, B, and C and Database 2 contains raw data for categories X, Y, and C. By identifying common data points in category C between databases 1 and 2, it becomes possible to make further connections among categories A, B, X, and Y. From here, we could identify common data points in a third database that contains raw data for categories E, F, and X. We believe that relational databases provide the best platform to capture this "web of relations" in big data. Quamen and Bath describe relational databases as "a series of interconnected spreadsheets. Each spreadsheet--called a table in database lingo--contains information on a real world entity such as People or Books or Songs or Birds or Rock Concerts or Places. Those tables are then tied together via relationships" (Quamen and Bath, 2016; 146-147). By providing a set of stepping stones or crosswalks between diverse datasets, relational databases enable us to build links between allied projects (i.e., ones that share a common data point) and more remote projects (i.e., ones that do not share a common data point) in order to combine expertise and mobilize already curated data in new environments.

Past Work

In 2014, MoEML research assistant Tye Landels-Gruenewald undertook a directed study course with director Janelle Jenstad with the aim of geocoding the *English Short Title Catalogue* (ESTC) from 1475 through 1666. With the generous assistance of David Eichmann and Blaine Greteman of the Shakeosphere project (based at the University of Iowa), we were able to extract toponyms from transcribed imprints in the ESTC catalogue using natural language processing (NLP) technology. We had intended on using named entity recognition to find matches between the extracted ESTC toponyms and our own gazetteer of early modern London locations; however, the toponyms themselves included too many errors or extra text to make this feasible. As Grover, Givon, Tobin, and Ball note in their white paper on "Named Entity Recognition for Digitised Historical Texts," there is still much work be done in order to teach named entity recognition software to recognize early modern English (Grover et al., 2008).

Concomitantly, Jenstad was manually compiling a spreadsheet of Pantzer numbers and cross-referencing them to MoEML location identifiers. Pantzer numbers are an alphanumeric string consisting of a letter and an integer. The letter indicates a general location. All the Pantzer numbers beginning with the letter O indicate locations in, near, or "against" the Royal Exchange. The numbers offer more granularity. For example, O.2 designates a location "at the north side of the Royal Exchange." Key challenges in matching Pantzer numbers with MoEML IDs were (1) different controlled vocabularies, and (2) the different levels of granularity inherent in the projects. Pantzer's authority names came from the imprint line wording; MoEML authority names are standardized spellings of the official or most common toponym variant (determined by set of critical rules we codified in order to build our gazetteer). Granularity differences emerged from the different interests of the two projects. Book historians map the bookstalls in the Royal Exchange, a location for which MoEML considered as a single entity (ROYA1); MoEML finer granularity emerges elsewhere, in our mapping of conduits, landings, and the many other precise locations that John Stow mentions in his Survey of London. A full crosswalk between Pantzer and MoEML would require the addition of sublocations to MoEML's placeography, a goal we will likely realize via the development of MoEML microsites for the Royal Exchange and Paul's Churchyard. In the meantime, we lose some of the granularity of Pantzer's data by assigning the same MoEML id to two or more Pantzer numbers.

Current Work

These past-attempts at establishing interoperability between datasets illustrate the challenges in attempting to traverse projects that only weakly share common data points. Between MoEML and the ESTC are a number of assumptions, potential errors, and remediations that weaken the link between the two respective datasets. To get to our larger project of mapping the STC, we must take smaller steps.

Our current work relates the playbook data collected by Zachary Lesser and Alan B. Farmer at *The Database of Early English Playbooks* (DEEP) to our own toponymic data, relying on Pantzer's vocabulary as a shared data-point. Jenstad's spreadsheet was transformed into a TEI-conformant XML table, which we ran across DEEP's openly available XML data. Doing so allows us to integrate DEEP numbers into the site, linking outwards to DEEP's newly static and predictable URLs.

The DEEP data and Pantzer-MoEML table can be related, but we recognize that this relationship is not immutable. In other words, both datasets are "living" databases insofar as the data can—and should be—curated and edited. Once Jenstad's spreadsheet was converted into TEI, Landels-Gruenewald was tasked with editing and refining Jenstad's initial findings to reflect the the growth of MoEML's gazetteer over the past two years (the MoEML team tagged nearly 2000 more toponyms between July 25, 2014– the last day Jenstad worked on the spreadsheet– and October 31, 2016, from 11,259 to 16,120). Lesser and Farmer have also recognized the need to amend Pantzer's findings in their data.

Future Work

The experiment with DEEP data has given us a stronger link to the ESTC. Now that we know Pantzer numbers are relatable to MoEML toponym IDs, we can now mobilize the data from Pantzer's appendix to connect MoEML with the ESTC. We plan to convert Pantzer's printed aggregations of STC numbers to digital files via OCR. With some curation, we will then have a list of all the STC numbers at each Pantzer number; using our crosswalk between Pantzer numbers and MoEML IDs, we will have a list of STC numbers (and therefore of unique print editions and issues) associated with MoEML locations. From the ESTC, we can obtain a crosswalk relating STC numbers to ESTC numbers. We add the caveat that Pantzer's locations will need to be corrected as book historians like Lesser and Farmer bring their knowledge to bear on her interpretation of STC data; every crosswalk dependent on her data will need to be refreshed and all the data maps remade. We can display these STC numbers as lists on MoEML location pages, much as Pantzer's print database does; in the digital environment, we can make dynamic links to DEEP or ESTC open-access pages for the book. We can also map these numbers on our open-layers Agas map platform as a layer of imprints associated with locations, eventually in combination with other tags (such as genre, now being added to EEBO by other scholars) or with other metadata fields harvested from the ESTC. All this data will pivot on the STC-MoEML data crosswalk that we are producing via Pantzer, following DEEP's initial work.

Distant Future Work

A longer-term goal is to harvest from the ESTC's XML files the strings of characters transcribed in the imprint line metadata field. Since we will already know from the STC-MoEML crosswalk which location is described in the imprint line, we can sort the imprint lines by locations and do rapid human scans for outliers, which may be a quick way of correcting Pantzer's data. We can also wrap TEI tags around the toponyms in the imprint lines, thereby increasing the number of toponymic variants in the MoEML gazetteer. The more variants in the gazetteer, the more accurate any future NER or geoparsing of large corpora will be. Given that we already search the EEBO-TCP corpus manually for references to place, we aspire to run our gazetteer against the entire TCP corpus to find and then map toponyms.

Principles and Practices of Curation for Future Mining and Interoperability

Acknowledging that the most interesting future uses of a project's data have not yet been imagined (Poston, 2011), how can we maximize the opportunities for other people to do things with that data? We suggest the following principles and practices as a starting point for discussion:

1. Make your data free to the world, preferably in easily downloadable and manipulable formats (in .json or .xml files, for example).
2. Be clear about how you compiled your data.
3. If you are aware of limitations in your data, tell the world.
4. As you correct and refine your data, communicate regularly about data updates.
5. If you are using other people's data in your own applications, check back regularly and rebuild the data crosswalks.
6. Know the weak link(s) in your data crosswalks.
7. Plan for corrections as other projects improve their data.

8. Be mindful of the potential for error to compound. Errors in my data, combined with errors in your data, have the potential to lead scholars to false conclusions.
9. Test your data crosswalks in a variety of ways. Take a small subset of the data and compare NLP results to hand curated results, for example.

Conclusion

Pantzer died in 2005, the year before MoEML was published at a public URL, but we like to think that she would have welcome the digital recreation, correction, curation, and connection of her data. She used the capacities of print to create a map and dense cross-references. Having "o'erleapt" Pantzer's curatorial work in building our digital catalogues, we now need to capture her formidable scholarship of interpreting and relating disparate types of data. We began with the goal of relating MoEML toponyms to ESTC numbers, but discovered that Pantzer's hand-curated data was more reliable than the results of NER and NLP. Our new question then became: "What sort of steps, processes, principles, and practices are necessary in doing this sort of work?" Handcrafted data, in conjunction with computer processing, allows for greater interoperability between projects and begins to achieve the possibilities of the data not conceived by Pantzer.

Bibliography

British Library Board. (n.d.) *English Short Title Catalogue (ESTC).* Available at: http://estc.bl.uk/ [Accessed 20 March 2017].

Farmer, A. and Lesser, Z., eds. (2007). *DEEP: Database of Early English Playbooks.* Available at: http://deep.sas.upenn.edu/ [Accessed 20 March 2017].

Farmer, A., and Lesser, Z. (2008). Early Modern Digital Scholarship and DEEP: Database of Early English Playbooks. *Literature Compass* 5(6), pp. 1139-1153.

Grover, C., Givon, S. Tobin, R., and Ball, J. (2008). Named Entity Recognition for Digitised Historical Texts. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Paris: European Language Resources Association, n. pag. Available at: http://www.ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf [Accessed 20 March 2017].

Jenstad, J., ed. *The Map of Early Modern London (MoEML).* Available at: http://mapoflondon.u vic.ca/ [Accessed 20 March 2017].

Pantzer, K. and Rider, P. (1991). *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland and of English Books Printed Abroad, 1475-1640.* Begun by A. Pollard and G. Redgrave. Vol. 3. London: Bibliographical Society.

Poston, M. (2011). *The most interesting use of our data will not be what we think it is.* [Blog] The Collation. Available at: http://collation.folger.edu/2011/12/the-most-interesting-use-of-our-data-will-not-be-what-we-think-it-is/ [Accessed 20 March 2017].

ProQuest LLC. (2003-) *EEBO: Early English Books Online.* Available at: http://eebo.chadwyck.com/home [Accessed 20 March 2017].

Quamen, H., and Bath, J. (2016). Databases. In: C. Crompton, R. Lane, and R. Siemens, eds., *Doing Digital Humanities: Practice, Training, Research.* London and New York: Routledge, pp. 145-162.

Stow, J. (1598). *A Survey of London.* London: Printed by John Windet for John Wolfe. STC 23341.

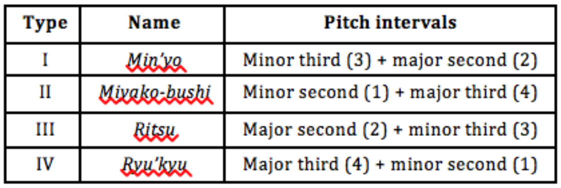
Regional Classification of Traditional Japanese Folk Songs from Southwest Regions

Akihiro Kawase
akawase@mail.doshisha.ac.jp
Doshisha University, Japan

Introduction

The main purpose of this study is to grasp the pitch transition patterns from pieces of traditional Japanese folk songs among southern three regions of Honshu (the Main Island of Japan), and by making comparisons with Koizumi's tetrachord theory in order to make the regional classification by the tendency in pitch information.

A tetrachord is a unit consisting of two stable outlining tones called kaku-on (nuclear tones), and one unstable intermediate tone located between the nuclear tone. Influenced by the methods of Western comparative musicology, the Japanese musicologist Fumio Koizumi conceived of a scale based not on the octave unit but rather on the interval of a perfect fourth, and has developed his tetrachord theory to account traditional Japanese music (Koizumi, 1958). Depending on the position of the intermediate tone, four different tetrachords can be formed (see Table 1 and Figure 1).



Type	Name	Pitch intervals
I	Min'yo	Minor third (3) + major second (2)
II	Miyako-bushi	Minor second (1) + major third (4)
III	Ritsu	Major second (2) + minor third (3)
IV	Ryu'kyu	Major third (4) + minor second (1)

Table 1. Koizumi's four basic tetrachords.

In the previous study, we have sampled and digitized the five largest song genres within the music corpora of the Nihon Min'yo Taikan (Anthology of Japanese Folk Songs, 1944-1993) from 45 Japanese prefectures, and have clarified the following three points by extracting and comparing their respective musical patterns (Kawase and Tokosumi 2011): (1) the most important characteristics in the melody of Japanese folk songs is the transition pattern, which is based on an interval of perfect fourth pitch that constructs Koizumi's four basic tetrachords; (2) regionally adjacent areas tend to have similar musical characteristics;

and (3) the differences in the musical characteristics almost match the East-West division in the geolinguistics or in the folkloristics from a broader perspective. However, to conduct more detailed analysis in order to empirically clarify the structures by which music has spread and changed in traditional settlements, it is necessary to expand the data and make comparisons based on the old Japanese provinces (ancient administrative units that were used under the ritsuryo system before the modern prefecture system was established).

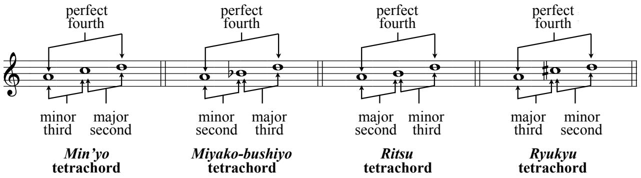


Figure 1. Example of four different types of Koizumi's tetrachord when taking A-D as kaku-on (nuclear tones).

Overview of data

We extracted the musical notes for works included in the "Nihon Min'yo Taikan", and digitized the entire scores from each province in the Kyushu district (geographically located in the southern part of Japan), Chugoku district (the westernmost region of Japan's largest island of Honshu), and Shikoku district (literally meaning four provinces, located south of Honshu and east of Kyushu district). In total, there were 474,191 tones in the sample of 2,383 songs for the 25 provinces (see Figure 2).

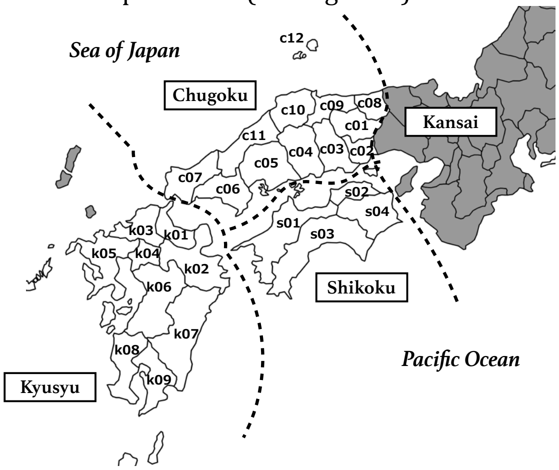


Figure 2. Geographical divisions of the three districts under the old provinces

Procedure

The procedures are as follows: (1) we digitized all the songs from each district and generated sequences that contain interval information from the song melodies; (2) extracted four transition probabilities of tetrachords for every province separately, and create a 24-dimensional data with 25 samples (provinces); and (3) applied principal components analysis (PCA) to identify patterns in the data, and to highlight their similarities and differences.

In order to digitize the Japanese folk song pieces, we generate a sequence of notes by converting the music score into MusicXML file format. We devised a method of digitizing each note in terms of its relative pitch by subtracting the next pitch height for a given MusicXML. It is possible to generate a sequence T that carries information about the pitch to the next note: T = (t1, t2, ... , ti, ..., tn). An example of the corresponding pitch intervals for ti can be written as shown in Table 2. We treat sequence T as a categorical time series, and execute N-gram to capture transitions and their trends.

Using a bigram model representing pitch transitions, all four types of tetrachords from Table 1 can be expressed as follows in ascending order: min'yo (+3, +2), miyako bushi (+1, +4), ritsu (+2, +3), and ryukyu (+4, +1). Depending on the positions of the three initial pitches in a tetrachord, six transition patterns can be considered in perceiving a tetrachord in two steps (bigram). Therefore, the amount of tetrachords within two steps can be obtained by counting the pairs of 24 transition patterns in sequence T.

t_i	Pitch Intervals	t_i	Pitch Intervals
0	perfect unison	7	perfect fifth
1	minor second	8	minor sixth
2	major second	9	major sixth
3	minor third	10	minor seventh
4	major third	11	major seventh
5	perfect fourth	12	perfect octave
6	aug.fourth/dim.fifth	13	minor ninth

Table 2. Corresponding pitch intervals

Results and discussions

The relative frequency of the first transition (unigram), is maintained between "-5" and "+5"; the interval of a perfect fourth pitch (see Figure 3). As the graph forms a symmetric shape with respect to "0", the pitch transitions occur almost equally in both descending and ascending order.

The implementation of the PCA is summarized as follows. As shown in Table 3, the component loadings for the first three principal components of each province explain more than 85.88% of the variability.

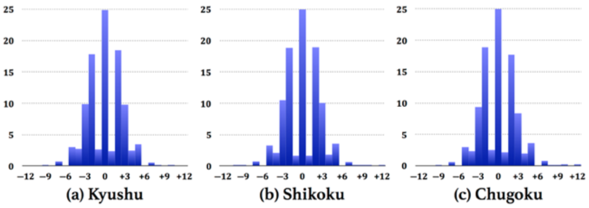


Figure 3. First transition frequencies for three regions

In the first column, the values of all 24 variables represent a positive quantity, and have almost the same weight. This result indicates that the profile of the first PCA axis is the persuasiveness of the tetrachord. Thus, as the value increases, the inclination of a pitch transition enhances its persuasiveness, and the value decreases, it loses its persuasiveness. In the second column, all variables for the min'yo and the ritsu represent a negative quantity, while 11 variables for the miyako-bushi and the ryukyu represent a positive quantity. According to ethnomusicological research, the min'yo and ritsu tetrachords appear frequently in Japanese folk songs. In contrast, the miyako-bushi and ryukyu tetrachords, steadily shifted from the ritsu and min-yo tetrachords respectively, and then increased in popularity as an emotional crutch (Koizumi, 1977). This result indicates that the profile of the second PCA axis is the relative pitch intervals between the nuclear tone and its intermediate tone, or in other words, the differences in patterns of transition from the nuclear tone. Thus, as the value increases, the adjacent intermediate tone forming the tetrachord tends to form a minor second interval (sort of a minor key progression), and as the value decreases, it tends to form a major second interval (sort of a major key progression).

The corresponding scores for each sample are plotted in a two-dimensional space to complete the PCA (Figure 4). We see that there is a strong contrast between min'yo, ritsu, miyako-bushi, and ryukyu. It is possible to clarify the structural commonalities and differences between areas. Figure 5 is the result of applying the hierarchical cluster analysis (Euclid distance, Ward method) to the corresponding scores. In addition, if we look for a height where there are three vertical lines and trace the lines down to the individuals, the partition corresponding to three clusters. If we plot this result on a map, we see that provinces are clearly classified according to geographical factors and cultural backgrounds (Figure 6).

Variables		1st PCA axis	2nd PCA axis	3rd PCA axis
minyo-1	(+3, +2)	0.947	-0.224	-0.021
minyo-2	(+5, -2)	0.893	-0.330	-0.136
minyo-3	(+2, -5)	0.879	-0.340	-0.136
minyo-4	(-3, +5)	0.888	-0.204	-0.151
minyo-5	(-2, -3)	0.951	-0.229	-0.090
minyo-6	(-5, +3)	0.876	-0.321	0.006
miyako-1	(+1, +4)	0.819	0.404	-0.107
miyako-2	(+5, -4)	0.684	0.531	-0.241
miyako-3	(+4, -5)	0.638	0.406	0.046
miyako-4	(-1, +5)	0.719	0.556	-0.170
miyako-5	(-4, -1)	0.779	0.552	-0.220
miyako-6	(-5, +1)	0.796	0.372	-0.103
ritsu-1	(+2. +3)	0.932	-0.267	0.006
ritsu-2	(+5, -3)	0.879	-0.298	-0.111
ritsu-3	(+3, -5)	0.797	-0.547	0.098
ritsu-4	(-2, +5)	0.904	-0.199	-0.190
ritsu-5	(-3, -2)	0.947	-0.244	-0.085
ritsu-6	(-5, +2)	0.784	-0.534	0.068
ryukyu-1	(+4, +1)	0.788	0.475	0.008
ryukyu-2	(+5, -1)	0.451	0.015	0.779
ryukyu-3	(+1, -5)	0.687	0.346	0.384
ryukyu-4	(-4, +5)	0.670	-0.067	0.508
ryukyu-5	(-1, -4)	0.699	0.606	-0.061
ryukyu-6	(-5, +4)	0.612	0.307	0.601
Eigen value		15.448	3.490	1.675
Percentage of var.		64.367	14.541	6.979
Cum.percentage of var.		64.367	78.909	85.888

Table 3. Component loadings for 24 transition patterns for tetrachords

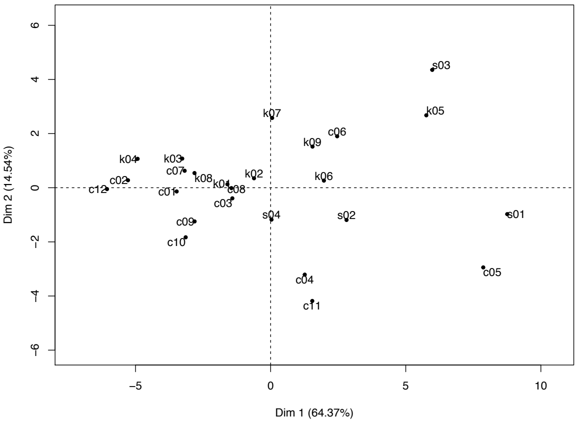


Figure 4. Plot of the first two component scores

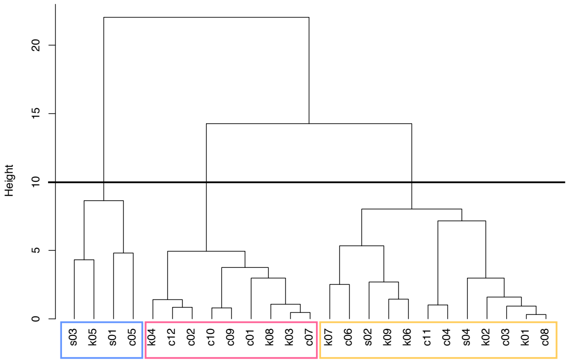


Figure 5. Dendrogram based on component scores using Ward method

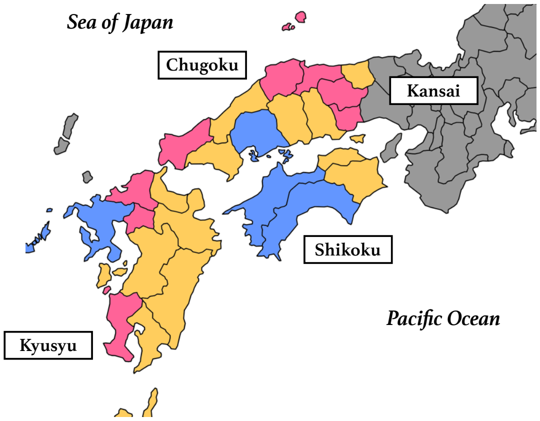


Figure 6. Clustering results plotted on a map

Conclusion

In this study, we digitized the melodies of endangered traditional Japanese folk songs from three regions, and quantitatively classified them according to the old province by executing principal components analysis and hierarchical cluster analysis in terms of pitch transitions based on a unit of Koizumi's tetrachord theory. As a result, compared to our previous studies on the small amount of data (e.g. Kawase 2016a, 2016b, 2016c), regions were successively classified according to both geographical factors and cultural backgrounds in detail, and classified the melodies into two basic groups according to the behavior of the intermediate tone. We firmly assured that the melodic structures of tetrachords in each province are shared by land and sea routes based on actual music data analysis.

Acknowledgments

This work was mainly supported by the Japanese Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (15K21601).

Bibliography

Kawase, A. and Tokosumi, A. (2011) Regional classification traditional Japanese folk songs, International Journal of Affective Engineering 10 (1): 19-27.

Kawase, A. (2016a) Extracting the musical schemas of traditional Japanese folk songs from Kyushu district, Proceedings of the International Conference on Music and Perception 14.

Kawase, A. (2016b) Regional classification of traditional Japanese folk songs from Chugoku district, Proceedings of the Annual Conference of Association for Digital Humanities 2016.

Kawase, A. (2016c) Melodic structure analysis of traditional Japanese folk songs from Shikoku district, Proceedings of the Sixth Annual Conference of Japanese Association for Digital Humanities 2016.

Koizumi, F. (1958) Studies on Traditional Music of Japan 1, Ongaku no tomosha.

Koizumi, F. (1977) Sounds in Japan, Seido-sha.

Kyoaki, N. H. (1944-1993) Nihon Min'yo Taikan (Anthology of Japanese Folk Songs), NHK Publishing.

MusicXML (n.d.) http://www.musicxml.com/for-developers/ [accessed 29 October 2016].

Quotidian Reading: Digitally Mapping Literary and Personal Geographies

Tom Keegan
thomas-keegan@uiowa.edu
University of Iowa, United States of America

Sarah Bond
sarah-bond@uiowa.edu
University of Iowa, United States of America

Students (and some teachers) feel rightly intimidated by Petronius' *Satyricon* or James Joyce's *Ulysses*. They are big books that are too often cast as *things* to be conquered or "done"— as in " I did *Ulysses*"—rather than encountered as portals to better understanding ourselves and the world in which we live. In the face of literary complexity and historical heft, students and instructors experience an anxiety of knowing. They feel expected to master a text rather than adapt it to their own needs or desires. In this long paper, we (Professors Sarah Bond and Tom Keegan) offer an alternate approach to reading texts in which the experiential learning advocated for by John Dewey (and often averred by literary theorists) is combined with a host of digital mapping tools, broadly understood. We describe our work in two courses—one in Classics and one in English—as aimed at connecting the content of Petronius' and Joyce's novels with the daily lives of our students. In our courses students undertook a kind of "quotidian reading" in which they identified spaces and practices in the novels and relocated those elements in their own lives, sharing their observations through mapping, blogging, and podcasting. The resulting coursework reversed the emphasis on knowing the text as thing, and placed it, instead, on knowing the text as a mode of self-understanding. Given both Joyce's and Petronius' preoccupation with space and geography, we invited students to map their own daily travels, applying the text as epically or mundanely as they saw fit. We were interested in how students came to see the text as relevant to them as everyday people.

This student-driven cartography in which real spaces are interpreted alongside "imaginary" spaces via digital means encourages humanities students to close the gap between what they are reading and what they are living. Quotidian reading has the potential to refigure all types of textual spaces—from the *Aeneid*'s description of the underworld to Dante's description of it over a millennium later—by inviting students to annotate and augment their spatial dynamics with textual references—and vice versa. In this paper, we lay out a set of best practices, digital methods, and core approaches to allowing students to thread the "imagined spaces" of the text into the daily spaces of our students' lives. Digital tools offer intriguing possibilities for reframing the stoic and epicurean spaces embedded in the *Satyricon* or resituating the daily environments and existential perspectives of *Ulysses'* Bloom, Stephen, and Molly. As our students mapping (among other things) the use, disuse, and misuse of public spaces in Iowa City, they approach a new understanding of Petronius' satirical spaces or Joyce's novel articulation of the public sphere. Through the use of the open-source tool "ImageMapster," which uses HTML5 canvases rather than Java in order to work in modern browsers, students represent, annotate, layer, and ultimately archive these literary topographies. The process allows students to engage in memoir through the geographic medium of the map and to remediate their experience of their daily life—in much the same way that Petronius and Joyce did. One result, has been the creation of personal "base maps" for the literary texts as constructed by our students. This product of the pedagogical enterprise suggests a new approach for commonly taught literary texts in other courses. Standards for the mapping of literary geographies as personal geographies are virtually non-existent, though they are a means for both engagement and reflection.

In some ways, this paper moves the application of GPS beyond Bill Clinton's 1996 desire to "encourage acceptance and integration of GPS into peaceful civil, commercial and scientific applications worldwide.” That broad public intent, combined with the pedagogical legacy of the “spatial turn” in literature, history, and art, has yielded a growing interest in digital humanities projects that make use of mapping as a tool of self-discovery. The pedagogical approach detailed in this paper expands on spatial turn’s address of the “microcosms of everyday life and the macrocosms of global flows.” And while cartography provides the ability to relate points spatially as they exist on the Earth’s surface, literature often deviates from the physical geometries denoted through longitude, latitude, and elevations coordinates. Through a quotidian reading, the spatial structure and topographical features found within novels such as Petronius’ *Satyricon* or James Joyce’s *Ulysses* help students organize, access, and navigate literary texts and themselves in daily life. The frequent tensions between the “real” and the “imaginary” geographies of a text and of our everyday lives can—when reflected upon—serve to unveil to students the emotions, perceptions, and motives of a character or author. This paper, we hope, will invite thoughtful conversation about the nature of the digital humanities in and beyond the classroom, as students are encouraged to see digital tools as ones they can use in everyday life.

Did a Poet with Donkey Ears Write the Oldest Anthem in the World? Ideological Implications of the Computational Attribution of the Dutch National Anthem to Petrus Dathenus

Mike Kestemont
mike.kestemont@gmail.com
University of Antwerp, Belgium

Els Stronks
e.stronks@uu.nl
Utrecht University, the Netherlands

Martine de Bruin
martine.de.bruin@meertens.knaw.nl
Meertens Institute, the Netherlands

Tim de Winkel
t.dewinkel@uu.nl
Utrecht University, the Netherlands

Introduction

The *Wilhelmus* has been the official national anthem of the Kingdom of the Netherlands since 1932. The song carries a wider relevance that extends well beyond the Low Countries. According to the authoritative *Guinness Book of Records*, the *Wilhelmus* is the national anthem with the oldest music in the world: we are able to date the tune and text to the years 1568-1572 during the Dutch Revolt, a key episode in the history of the Early Low Countries. Moreover, in the song’s fifteen couplets, an anonymous poet has immortalized a dramatic internal monologue of William the Silent, Prince of Orange (1533 – 1584), a well-known figure who has played a decisive role in the political history of Europe (Van Stipriaan, 2007).

In the earliest sources, the *Wilhelmus* has invariably survived anonymously, in print collections of rebel songs (the so-called *geuzenliederen* or ‘beggar songs’) that date back to the Spanish Occupation in the Low Countries (De Bruin, 1998). Only some of these songs are attributed to known authors; the majority, including the *Wilhelmus*, are not. Apart from the supposed date of composition (1568-1572), there are few historical facts that could help the attribution. Although the *Wilhelmus* does not explicitly choose sides in contemporary religious conflicts, circumstantial evidence strongly suggests that the text was written by an author of Flemish or Dutch descent, who was living in a German refugee community at the time, perhaps in the vicinity of Heidelberg, because of a number of striking intertextual connections to other songs that were composed in that area.

Ever since its creation in the late sixteenth century, the attribution of the song has not ceased to puzzle scholars as well as other inhabitants of the Low Countries. Only decades after the song’s composition, there seems to have been considerable confusion already: in various sources, we find widely divergent attributions of the song to a number of famous authors, such as Marnix of Saint Aldegonde (the mayor of Antwerp, during the city’s famous Fall in 1585) or the religious author and philosopher Dirck Coornhert. Many other candidate authors would be suggested in the next centuries, the credibility of which could vary strongly. In the public opinion, Marnix has long remained the most popular candidate, although scholars have never reached any definitive agreement on the issue. As late as 1996, for instance, an entire doctoral thesis was devoted to the authorship of the *Wilhelmus*. In this thesis, Maljaars predominantly argued that Marnix could not have been the author, relying on traditional evidence: the results of the close reading of the Wilhelmus, and comparison between the *Wilhelmus* and other texts by the presumed author(s).

In 2016 an interdisciplinary team of scholars has tackled this age-old issue from a new perspective: stylometry. For most of the candidate authors which have been suggested for the *Wilhelmus*, we have available relatively sizable oeuvres of lyrical poems or even highly similar songs. The comparison of the *Wilhelmus* to those reference oeuvres, using state of the art stylometric methodologies, should allow us to estimate the relative distance from the anthem to each candidate author (*authorship attribution*) and verify their authorship (*authorship verification*). Many issues, however, make this comparison far from trivial: the texts are short (the *Wilhelmus* only counts 500 words), we only know younger, potentially corrupted versions of the texts and rarely have autographs, the spelling of the material is highly unstable etc. We have tried to tackle the latter issue through part-of-speech tagging and lemmatizing the texts (Kestemont et al., 2016b): instead of performing measurements on the original surface forms, we would restrict our analyses to the most frequent tag-lemma pairs (MFTLPs), which normalize the spelling of tokens.

In this paper, we will report several authorship experiments, using both the attribution and the verification setup (Kestemont et al., 2016a), in which we have compared the *Wilhelmus* to a representative set of contemporary authors, among which the main candidate authors as well as some background authors that merely served as 'distractors' or 'imposters'. We include a small selection of these below. Surprisingly, these experiments without exception pointed towards an obscure, vilified author who has never even been mentioned as a candidate author: Petrus Dathenus (ca. 1531-1588). The first series of plots are rather naive Principal Components analyses (300 MFTLPs) which each confront textual samples by two candidate authors and the *Wilhelmus* (in white). In these binary comparisons, the *Wilhelmus* is attributed to Dathenus without exception. The same goes for the verification experiment (which runs entirely parallel to the experiments run on the Caesarian corpus in Kestemont, et al. 2016a): when compared to both target and imposter authors, the *Wilhelmus* is significantly closer to Dathenus's texts than to any other candidate author from this period for which we have texts available.

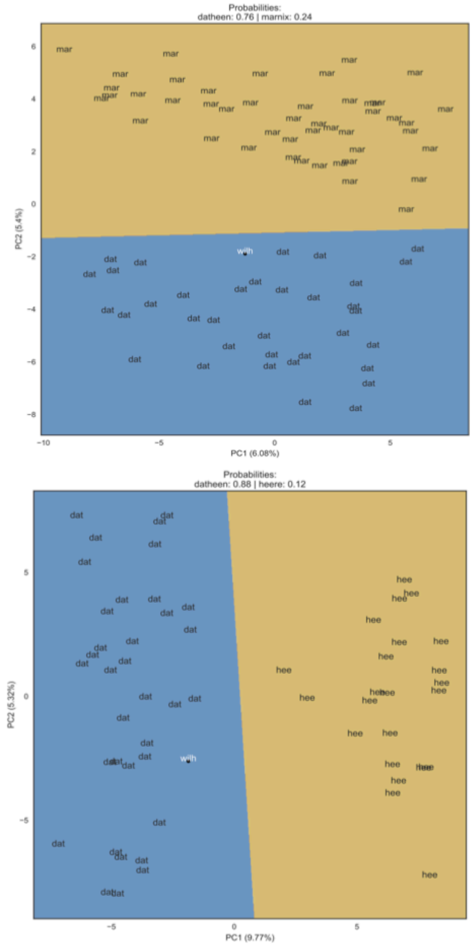


Fig. 1: Naive 2-dimensional PCA plots in which textual samples by two authors (Datheen vs Marnix; Datheen vs Heere) are confronted, including the *Wilhelmus*.

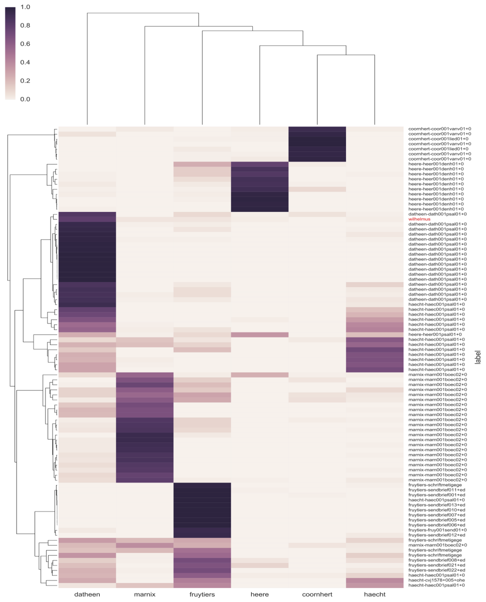


Fig. 2: Cluster map for the verification results obtained for the *Wilhelmus* and a number of highly relevant candidate authors (Kestemont et al. 2016a).

Dathenus is primarily known as the author of a complete Dutch adaptation of the Psalms, which became extremely influential in the second half of the sixteenth (and which is in fact still sung today in some reformed communities). His contemporaries considered him a great and dangerous orator because of his convincing way with words. Nowadays, Datheen has the reputation of being a very poor poet. To what does he owe this bad reputation? Our present-day image of the man goes back to the late eighteenth century when the pressure grew to have his Psalm adaptation replaced by a more modern one in churches. In order to increase the pressure on Dathenus's Psalms, people started mocking the poet through the dissemination of caricatures in which the man would even be depicted with donkey ears (see Fig. 3). It is striking how strongly our present-day view of Dathenus is still determined by the highly anachronistic eighteenth image of this author, instead of that of the respected and influential individual he was known to be in his own time.

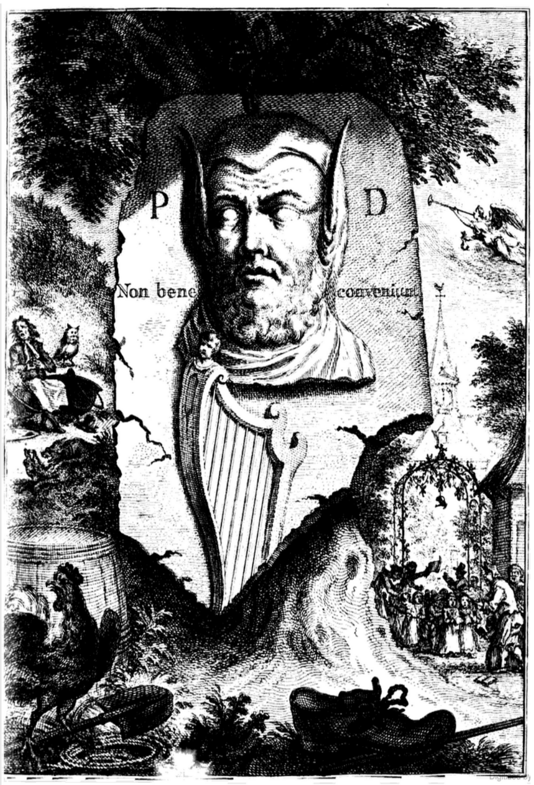


Fig. 3: A late eighteenth century caricatural depiction of Petrus Dathenus with donkey ears, to symbolize his alleged poetical ignorance.

In this paper, we will re-assess the sparse, historical evidence that is available for Petrus Dathenus and show that he is, in fact, an unusually strong authorial candidate for the *Wilhelmus*. Here, we limit our discussion to a single new fact that recently emerged. The *Wilhelmus* is a so-called contrafact: the song has been composed by writing a new set of lyrics for an already existing melody, a very common practice in early modern song culture. The original melody which was used for the *Wilhelmus* was a French song: *O la folle entreprise du Prince de Condé*. Musicologist have been able to pinpoint when this song was created: it must have been composed (as a Protestant song) during the Siege of Chartres in 1568. The tune must have been introduced in the Low Countries via the *Wilhelmus* and was not known beforehand. Therefore, it has always puzzled scholars how the *Wilhelmus* author might have been exposed to this French tune. Intriguingly, it turns out that Dathenus must have been present at the Siege of Chartres as a field preacher on the protestant side. Thus, although he has never made it to the official candidate list, Dathenus is in fact the only candidate, who not only has the right stylistic profile, but of whom we also argue that he was directly exposed to the base tune of the *Wilhelmus*.

In our paper, we will not go as far as to claim that the neglect of Petrus Dathenus as a potential candidate author for the national anthem of the Netherlands has been an ideological 'cover up operation'. We will discuss, however, the anachronistic biases and prejudices which so far have prevented the identification of Petrus Dathenus as a potential candidate author. From the point of Digital Humanities, it is important to stress that we base this research on a bold computational attribution to an author who, at first sight, seems a highly unlikely candidate; a human expert would never even have dared to think of this attribution. Nevertheless, exactly because machines do not carry the same set of preconceptions as humans, the application of stylometry is able to induce serendipity in humanities research and open up new perspectives.

Bibliography

De Bruin, M. (1998). 'Het Wilhelmus tijdens de Republiek'. In: *Volkskundig Bulletin* 24, p. 16-22.

Maljaars, A. (1996). *Het Wilhelmus: auteurschap, datering en strekking: een kritische toetsing en nieuwe interpretatie*. Kampen.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, K. and Daelemans, W. (2016a) 'Authenticating the writings of Julius Caesar'. In: *Expert Systems with Applications* 63: pp. 86–96.

Kestemont, M.; De Pauw, G.; Van Nie, R. and Daelemans, W. (2016b). Lemmatization for variation-rich languages using deep learning. In: *Digital Scholarship in the Humanities*, advanced access: http://dx.doi.org/10.1093/llc/fqw034.

Ruys, T. (1919). *Petrus Datheen*, Amsterdam.

Van Stipriaan, R. (2007). 'Words at War. The Early Years of William of Orange's Propaganda.' In: *Early Modern History* 11, p. 331-349.

Script Identification in Medieval Latin Manuscripts Using Convolutional Neural Networks

Mike Kestemont
mike.kestemont@gmail.com
University of Antwerp, Belgium

Dominique Stutzmann
dominique.stutzmann@irht.cnrs.fr
Centre National de la Recherche Scientifique, France

Introduction

In paleography, scholars study the history of handwriting, a crucial aspect of book history and manuscript studies. Paleography has traditionally been dominated by expert-based approaches, driven by the opinions of a small group of highly trained individuals. These have acquired a set of expert skills through year-long training, e.g. the ability to date a handwriting or attribute it to specific individuals. This knowledge remains very hard to render explicit, in order to share it with others. Therefore, paleographers are increasingly interested in digital modelling techniques to enhance the creation and dissemination of paleographic knowledge (Stutzmann, 2015). An important task in paleography is the classification of script types, especially now that digital libraries (BVMM, Gallica, e-Codices, Manuscripta Mediaevalia, etc.) are amassing reproductions of medieval manuscripts, often with scarce metadata. Being able to recognize the script type of such historic artefacts is crucial to date, localize or (semi-)automatically transcribe them. This paper focuses on script identification for medieval Latin manuscripts (ca. 500 AD to 1600 AD) and demonstrates the feasibility of a fairly accurate, meaningful automated classification.

Medieval script classification was the focus of the recent CLaMM (Classification of Latin Medieval Manuscripts) competition. For this shared task, the organizers released a training data set of 2,000 photographic (greyscale, 300 dpi) reproductions of pages extracted from Latin manuscripts, which were classified into a 12 script type classes, including uncial, caroline, textualis and humanistic script, but also more difficult to delineate classes such as the cursiva or (semi)hybrida. The participating teams had to submit a standalone application which was able to classify unseen images and estimate the distance between them. The organizers would then apply the submissions to 1,000 resp. 2,000 test images (Stutzmann, 2016) and evaluate their performance using various evaluation schemes. Here, we discuss the *DeepScript* submission to the CLaMM competition. The competition's results have officially been released on 26 Oct. 2016. *DeepScript* was ranked first on task 2, i.e. the 'crisp' classification of mixed script images (Cloppet et al., 2016). As the ground truth and results were released too recently, we limit this abstract to a general discussion of the approach; the final version and presentation of this paper will be supplemented with additional information and test results.

The *DeepScript* submission builds upon recent advances in Computer Vision, where the use of so-called 'deep' neural networks has recently led to dramatic breakthroughs in the state of the art of image classification (LeCun et al., 2015). The kind of neural networks applied in Computer Vision are typically convolutional: they slide small 'filters' (feature detectors) across images to make the network robust to small translations of objects. The networks make use of many 'layers' of such feature detectors, where the output of one feature detector always feeds into the next one. The use of such a stack of layers is beneficial, because this 'deep architecture' allows algorithms to model features of an increasing complexity (Bengio et al., 2013): in the first layers of the network, very raw and primitive shapes are detected ('edges'); it is only at the higher layers in the networks that these primitive features are combined into more complex, abstract visual patterns (e.g. entire faces). These neural networks lie at the basis of e.g. modern face verification algorithms on social media websites such as *Facebook*.

Neural networks are composed of millions of parameters which have be optimized. For this, the available training data is split out in a set of training images and a smaller set of development images (respectively ca. 1,800 and 200 images): the former is used to optimize the parameters of the network during training, the latter is used to monitor the performance of the network. The use of development data is necessary to avoid 'overfitting': it is possible for a network to start 'memorizing' the training images, so that it produces perfect predictions for the training data, but is not able any more to generalize properly to new, unseen images. By using a development set, we can stop optimizing the network, if its predictions for the development data do not increase in quality anymore. Only at this stage, the algorithm is evaluated on the actual test images.

Modern neural networks are typically trained on hundred thousands of training images. In the field of Cultural Heritage data, a common challenge is that most data sets are much smaller, and CLaMM is no exception, so that the danger of overfitting is much larger. We therefore proceeded as follows: the generous resolution for each training image was downsized by one half. Next, we would select random square crops or patches from the image (150x150 pixels) and train the algorithms on batches of these crops. This approach is blunt, yet innovative, since we make no effort to extract more specific regions of interest from the images, such as individual lines, words or characters. To avoid overfitting, we also applied augmentation: each training crop would be 'distorted' through randomly varying the zoom level, rotation and translation. Introducing such noise in the input is a common strategy to combat overfitting. Below goes an example of such a set of augmented patches for a single manuscript page (Fig. 1).

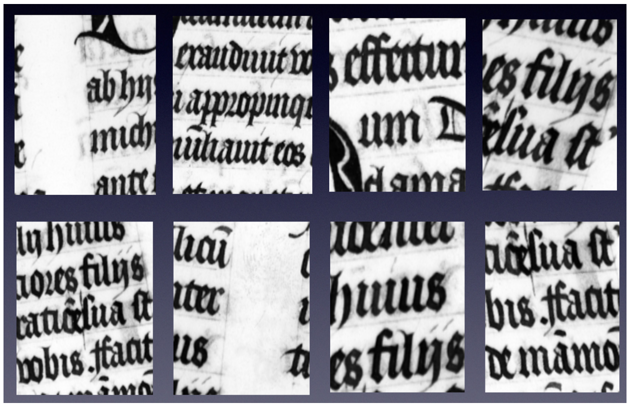


Fig. 1: Example of augmented crops for a single manuscript page.

After each epoch, we evaluated the performance of the current state of the network through inspecting the classification accuracy on the development images: we randomly selected 30 crops from each image (without augmentation), and calculated the average probability for each output class. The full image was assigned to the class with the highest average probability. The best validation accuracy which we achieved was 91.17%, using a network architecture of 14 layers, inspired by the famous Oxford VGG net (Simonyan et al., 2015). The manual classification of CLaMM images was based on morphological differences and allographs, as defined in standard works on Latin scripts such as (Bischoff, 1986; Derolez, 2003). The confusion matrices in Fig. 2-4 for the actual and the predicted classes in the development and test data illustrate that the confusions generally make sense from a paleographic point of view (the normal textualis letter is for instance often mistaken for the Southern or semi-textualis variant).

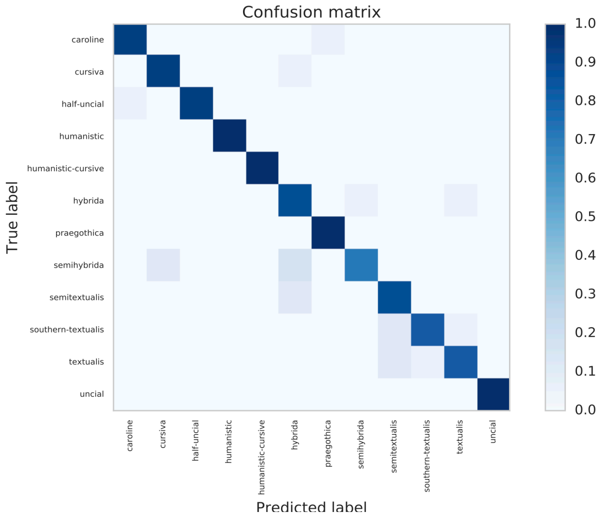
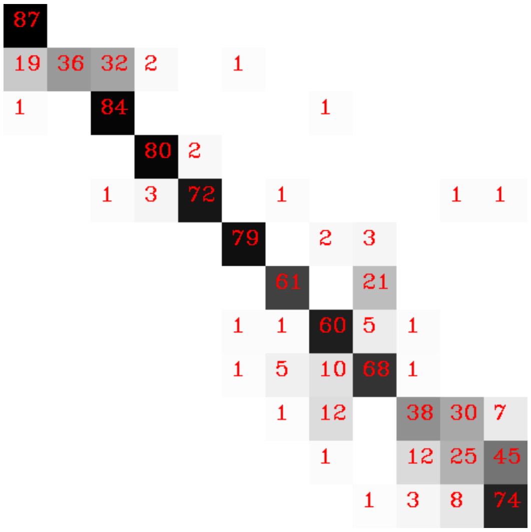
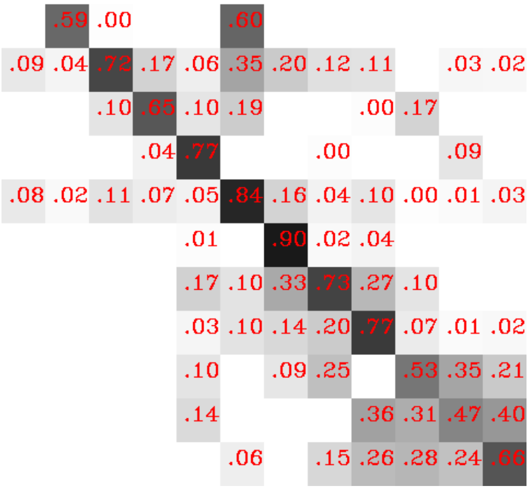


Fig. 2: Confusion matrix for the development data.



87											
19	36	32	2		1						
1		84					1				
			80	2							
		1	3	72		1				1	1
					79		2	3			
						61		21			
					1	1	60	5	1		
					1	5	10	68	1		
						1	12		38	30	7
							1		12	25	45
								1	3	8	74

Fig. 3: Classifications for the test data as a confusion matrix (task 1). Horizontal lines: ground-truth; Vertical columns: predictions. Order: 1-Uncial; 2-Half-uncial; 3-Caroline; 4-Humanistic; 5-Humanistic; Cursive; 6-Praegothica; 7-Southern Textualis; 8-Semitextualis 9-Textualis; 10-Hybrida 11-Semihybrida 12-Cursiva.



	.59	.00			.60						
.09	.04	.72	.17	.06	.35	.20	.12	.11		.03	.02
		.10	.65	.10	.19			.00	.17		
			.04	.77			.00			.09	
.08	.02	.11	.07	.05	.84	.16	.04	.10	.00	.01	.03
				.01		.90	.02	.04			
				.17	.10	.33	.73	.27	.10		
				.03	.10	.14	.20	.77	.07	.01	.02
				.10		.09	.25		.53	.35	.21
				.14				.36	.31	.47	.40
					.06		.15	.26	.28	.24	.66

Fig. 4: Membership Degree matrices for task 2, on the 999 test images where only one label is attributed to the image.

There exist interesting methods to visualize which patterns the trained network is sensitive to. Using the principle of gradient ascent, we start from a random noise image and feed it to one of the filters on the last convolutional layer: during 3,000 iterations we change the image so that it maximizes the activation of this particular filter. In Fig. 3, we show the 25 artificially generated images which yielded the strongest results; clearly, the network picks up relevant patterns. The third image from the left in the first line, for instance, clearly captures the presence of loops in the ascenders of individual characters (e.g. in the 'b' or 'h', which is crucial to differentiate between e.g. a textualis and a cursive letter). These visualizations directly tackle the issue of the computational 'black box' in the Digital Humanities, and espsecially Digital Palaeography (Hassner et al., 2013; Stutzmann et al., 2014). In our paper, we will offer further interpretations and visualizations of our model and confront these with the results from other participants in the CLaMM competition to offer new perspectives on the graphic definition of script classes in traditional paleography.

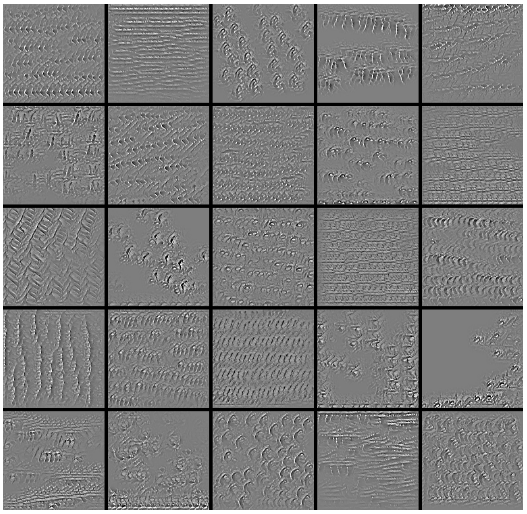


Fig. 5: Artificially generated images that maximally activate filters in the final convolutional layer.

Bibliography

Bengio, Y., Courville, A. and Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *TPAMI* **35**:8, 1798–1828.

Bischoff, B., *Paläographie des römischen Altertums und des abendländischen Mittelalters*, Berlin, 1986.

Cloppet, F., Eglin, V., Kieu, V. C., Stutzmann, D. and Vincent, N. (2016), ICFHR2016 Competition on the Classification of Medieval Handwritings in Latin Script, *Proceedings of the ICFHR 2016*, 590-595.

Derolez, A., *The Palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*, Cambridge, 2003.

Hassner, T., Rehbein, M., Stokes, P. and Wolf, L. 2013. Computation and palaeography: Potentials and limits. *Dagstuhl Manifestos* **2**: 14–35.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* **521**(7553): 436–44.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *Proceedings of ICLR 2015.* https://arxiv.org/abs/1409.1556/.

Stutzmann, D. and Tarte, S. (2014). Digital palaeography: New machines and old texts : Executive summary. *Dagstuhl Reports* 4.7: 112–34 (112–14, 132).

Stutzmann, D. (2015). Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol. *Digital Medievalist* **10**, http://digitalmedievalist.org/journal/10/stutzmann/.

Stutzmann, D. (2016). *ICFHR2016 Competition on the Classification of Medieval Handwritings in Latin Script.* http://icfhr2016-clamm.irht.cnrs.fr/.

What Changed When Andy Weir's *The Martian* Got Edited?

Erik Ketzan
eketza01@mail.bbk.ac.uk
Birkbeck, University of London, United Kingdom

Christof Schöch
christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Introduction

The Martian is a best-selling science fiction novel by Andy Weir that became a hit film in 2015. The novel exists in two versions, or variants: Weir self-published *The Martian* on his personal website in 2011 (hereafter, "*Martian1*") and began selling it on Amazon.com in 2012. Crown Publishing subsequently bought the rights, edited the book, and re-released it (hereafter, "*Martian2*").

The research presented here investigates what exactly changed when *The Martian* got edited. At first glance, the two versions appear essentially the same, with no major changes to plot, character, or structure. A closer look using a combination of quantitative and qualitative methods, however, reveals a number of noteworthy changes, as well as notable changes that result from thousands of seemingly minor copyedits.

Aims

The aim of our research is to identify what changed between the two variants of *The Martian* using a combination of close reading and digital methods, analyze why those changes are important, and propose a methodology for comparing self-published and later-edited novels, an increasingly common phenomenon. We hypothesize that the editing process of a leading publishing house results in a novel that is more "mainstream", i.e. socialised, domesticated, and appealing to a general audience. In order to test this hypothesis, we explore a range of aspects, including style, content, and character. Our research also aims to bring a critical perspective to the strengths and weaknesses of a variety of qualitative and technical methods in identifying the edits and assessing their importance.

Related Work

In addition to work in digital genetic criticism (e.g. van Hulle 2008), a small number of studies use digital methods to explore variants of contemporary fiction. Yufang Ho (2011) compared the 1966 and revised 1977 versions of John Fowles's novel *The Magus*, while Martin Paul Eve (2016) looked at differences in the US and UK versions of David Mitchell's *Cloud Atlas*. As both Ho and Eve use different methods from one another and from us, it appears that no standard method has emerged so far for this type of research.

Data

The data used for this research is primarily two plain text files of the variants of *The Martian*. *Martian1* was obtained in PDF format from Andy Weir's website. *Martian2* was obtained by scanning a print copy, performing OCR with manual corrections. We consider this our best option given the legal issues regarding text protected by copyright.

Methods and Results

Basic collation

We used the Wdiff frontend to the "diff" algorithm (Hunt & McIlroy 1975) to produce a collated version of *Martian1* and *Martian2* and assess the number and extent of the edits. We then used bespoke Python scripts to classify the edits identified by Wdiff.

We found a total of 5146 edits were made to the novel. While 92% of the 101,000 words in *Martian1* remain unchanged in *Martian2*, the remaining 8% of the words undergo some type of edit, whether they are deleted or modified (Figure 1). The sheer number of edits calls for automatic means to classify them and detect any patterns.

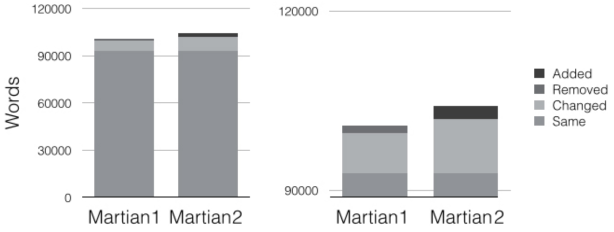


Figure 1: Visualization of edits to The Martian as grouped by Wdiff.

Automatic Classification of Edits

Edits were automatically classified into two broad categories: script-detectable copyedits, and all other edits. Script-detectable copyedits includes changes in capitalization, whitespace, hyphenation, spelling of numbers, abbreviations, or combinations thereof (Figure 2). All other edits were classified as insertion, deletion, expansion or condensation and as "minor" or "major", depending on the Levenshtein distance (Figure 3). Of the 5146 edits, 2863 (or 55%) were script-detectable copyedits, while 2283 (or 45%) comprised the rest. The code used as well as the collation data obtained are available on GitHub.

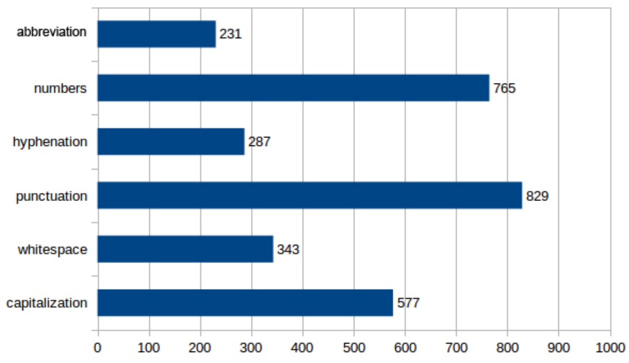


Figure 2: Script-identifiable copyedits to *The Martian.*

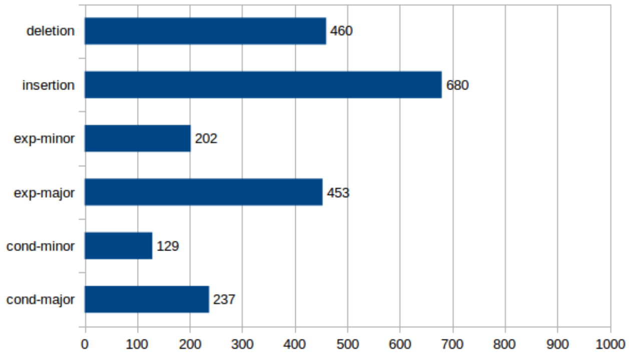
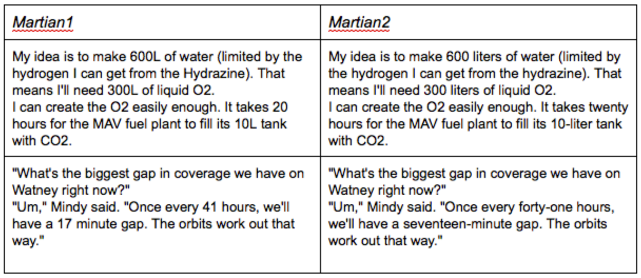


Figure 3: All other edits to *The Martian.*

Cumulative Effect of the Script-Identifiable Copyedits

Taken together, the 2863 script-identifiable copyedits have substantial effects upon the text. Weir's many misspellings and misuse of hyphens and capitalization are corrected. Numbers in *Martian1* are overwhelmingly written numerically, and 765 of these become words in *Martian2*, e.g. "8" becomes "eight". We found 231 instances of edits involving abbreviations, e.g. "L" becomes "liters".

The copyedits work together in different ways when they appear in protagonist Mark Watney's narration or in sections written in the third person (Figure 4). When Watney narrates, the hundreds of misspellings, numerals, and scientific abbreviations in *Martian1* support the fiction that he is a scientist working in extreme conditions. *Martian2* increases readability but eliminates the stylistic realism of Watney's text. When Weir uses, for instance, numerals in the dialogue of other characters, the effect can be jarring. *Martian2* corrects this for the better.



Martian1	*Martian2*
My idea is to make 600L of water (limited by the hydrogen I can get from the Hydrazine). That means I'll need 300L of liquid O2. I can create the O2 easily enough. It takes 20 hours for the MAV fuel plant to fill its 10L tank with CO2.	My idea is to make 600 liters of water (limited by the hydrogen I can get from the hydrazine). That means I'll need 300 liters of liquid O2. I can create the O2 easily enough. It takes twenty hours for the MAV fuel plant to fill its 10-liter tank with CO2.
"What's the biggest gap in coverage we have on Watney right now?" "Um," Mindy said. "Once every 41 hours, we'll have a 17 minute gap. The orbits work out that way."	"What's the biggest gap in coverage we have on Watney right now?" "Um," Mindy said. "Once every forty-one hours, we'll have a seventeen-minute gap. The orbits work out that way."

Figure 4: Edits to numerals and scientific abbreviations in Watney's narration (top) and third-person character dialogue (bottom).

Detecting transpositions with CollateX

Wdiff does not detect transpositions, or text that has been moved to a different location in the novel. Using CollateX (Dekker & Middell 2011) as described in Schöch (2016) revealed a total of 126 transpositions. Twenty-eight (or 22%) involve punctuation and should be considered artefacts of the method; 43 (or 34%) represent transpositions of a single word, showing stylistic preferences on the word-order level; 55 (or 44%) concern multi-word expressions which change the overall construction of a sentence or paragraph more substantially.

Figure 5 shows a relatively minor transposition appearing in combination with a contraction of a sentence.

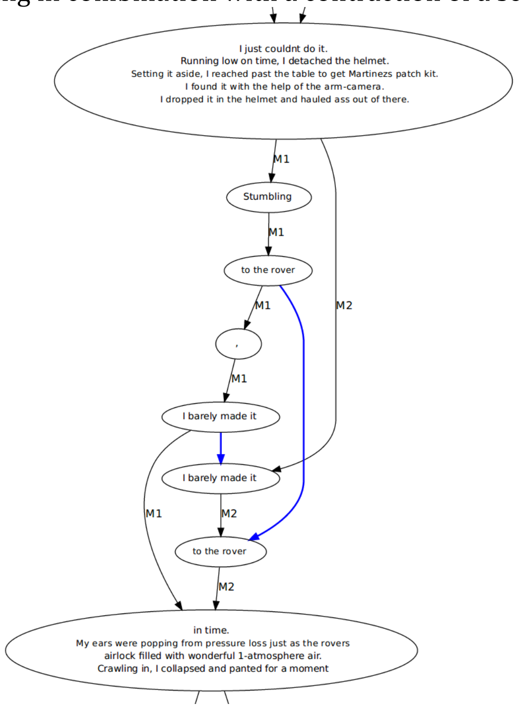

I just couldnt do it.
Running low on time, I detached the helmet.
Setting it aside, I reached past the table to get Martinezs patch kit.
I found it with the help of the arm-camera.
I dropped it in the helmet and hauled ass out of there.
M1
Stumbling
M1
to the rover
M1
M2
,
M1
I barely made it
I barely made it
M1
M2
to the rover
M2
in time.
My ears were popping from pressure loss just as the rovers
airlock filled with wonderful 1-atmosphere air.
Crawling in, I collapsed and panted for a moment

Figure 5: An example of a transposition identified by CollateX.

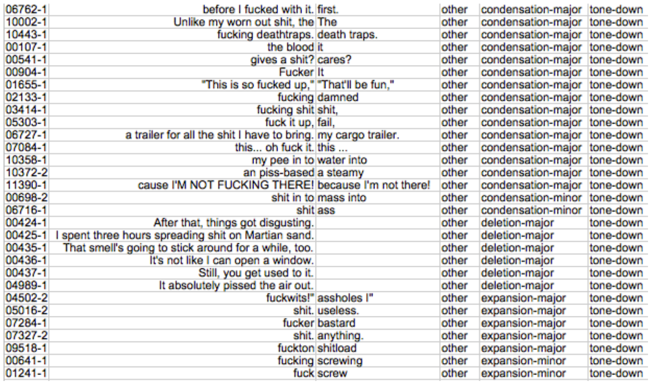
We conclude that, quantitatively and qualitatively, transpositions were not a major part of the edit to *The Martian*. However, future work could apply the same method to other, comparable variants of novels to gain better reference points.

Close Reading of Other Edits

When we grouped the other edits, placed them into a spreadsheet, and manually inspected them, a number of thematic and stylistic shifts between *Martian1* and *Martian2* became apparent.

Profanity is a key stylistic feature of *The Martian* that is substantially cut and softened by the edit. Words like “fuck” and “shit” are substantially reduced (by about 33% and 15%, respectively), while numerous other words and phrases are softened with “lesser” profanity or simple non-profanity (e.g. “the shit hits the fan” becomes “all hell breaks loose”). Figure 6 shows a selection of these edits. Similarly, crude and sophomoric humor is cut in key instances.

The plot of *The Martian* revolves around solving one problem after another to rescue an astronaut, Mark Watney, stranded on Mars, while relatively little text is devoted to Watney’s emotions or inner world. In *Martian2*, however, Watney expresses significantly more emotion: he misses his family and friends more and expresses despair, loneliness, and introspection more often.



06762-1	before I fucked with it.	first.	other	condensation-major	tone-down
10002-1	Unlike my worn out shit, the	The	other	condensation-major	tone-down
10443-1	fucking deathtraps.	death traps.	other	condensation-major	tone-down
00107-1	the blood	it	other	condensation-major	tone-down
00541-1	gives a shit?	cares?	other	condensation-major	tone-down
00904-1	Fucker	It	other	condensation-major	tone-down
01655-1	"This is so fucked up,"	"That'll be fun,"	other	condensation-major	tone-down
02133-1	fucking	damned	other	condensation-major	tone-down
03414-1	fucking shit	shit,	other	condensation-major	tone-down
05303-1	fuck it up,	fail,	other	condensation-major	tone-down
06727-1	a trailer for all the shit I have to bring.	my cargo trailer.	other	condensation-major	tone-down
07084-1	this... oh fuck it.	this ...	other	condensation-major	tone-down
10358-1	my pee in to	water into	other	condensation-major	tone-down
10372-2	an piss-based	a steamy	other	condensation-major	tone-down
11390-1	cause I'M NOT FUCKING THERE!	because I'm not there!	other	condensation-major	tone-down
00698-2	shit in to	mass into	other	condensation-minor	tone-down
06716-1	shit	ass	other	condensation-minor	tone-down
00424-1	After that, things got disgusting.		other	deletion-major	tone-down
00425-1	I spent three hours spreading shit on Martian sand.		other	deletion-major	tone-down
00435-1	That smell's going to stick around for a while, too.		other	deletion-major	tone-down
00436-1	It's not like I can open a window.		other	deletion-major	tone-down
00437-1	Still, you get used to it.		other	deletion-major	tone-down
04989-1	It absolutely pissed the air out.		other	deletion-major	tone-down
04502-2	fuckwits!"	assholes !"	other	expansion-major	tone-down
05016-2	shit.	useless.	other	expansion-major	tone-down
07284-1	fucker	bastard	other	expansion-major	tone-down
07327-2	shit.	anything.	other	expansion-major	tone-down
09518-1	fuckton	shitload	other	expansion-major	tone-down
00641-1	fucking	screwing	other	expansion-minor	tone-down
01241-1	fuck	screw	other	expansion-minor	tone-down

Figure 6: examples of toned-down profanity in the editing of The Martian

Additionally, *Martian1* contains an epilogue that is completely cut in the edit. It portrays Watney, back on Earth, being openly and profanely rude to a young fan. In *Martian2*, meanwhile, text is added to have Watney express gracious appreciation for all the parties involved in his rescue and a widespread faith in human nature. The edit therefore alters the tone of the ending substantially.

We believe that all of these changes, analyzed together with close reading, serve to align Watney’s character with our overall hypothesized goal of the edit: to make Watney more “relatable,” “nice,” and “human,” and thus to appeal to a wider audience.

Edits Over the Course of the Novel

Patterns in the edits related to textual progression are revealed by measuring the absolute Levenshtein distance of the script-identifiable copyedits and other edits line by line (Levenshtein distance is a metric for measuring the difference between two sequences, see Navarro 2001).

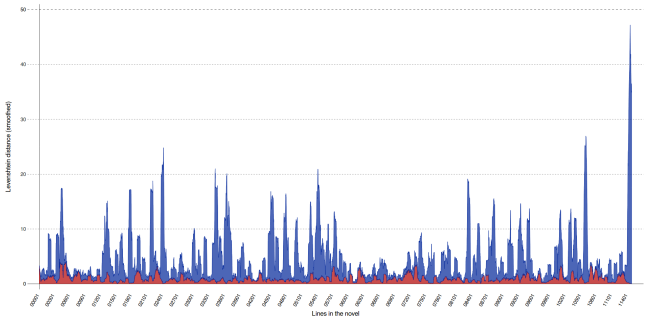


Figure 7: Sum of absolute Levenshtein distance per line over textual progression (script-identifiable copyedits in red, other edits in blue).

Figure 7 shows the sum of the absolute Levenshtein distances for each line of the novel (with Savitzky-Golay smoothing applied). The graph shows the substantial modifications to the ending of the novel, but also a large number of locations with smaller but nonetheless above-average modifications.

Conclusion and Further Research

We have identified and analyzed a number of key features that emerged from the editing of *The Martian*, notably on the level of style and character, which combine to make the novel more appealing to a wider audience.

Ongoing research into *The Martian* concerns the relative frequency and function of parts of speech, quantifying the amount of syntactic change, and the legal issues affecting the obtaining and processing of the texts. We hope to present these additional findings in the near future.

As for our typology of edits, an established methodology for classifying edits in the companion fields of textual analysis and scholarly editing is the distinction between the “accidentals” and “substantives” used by the Greg-Bowers tradition and included in the MLA Committee on Scholarly Editions’ *Guidelines for Editors of Scholarly Editions* (Modern Language Association, 2011). Scholars are not unanimous, however, in supporting this. G. Thomas Tanselle, for instance, found these terms “misleading and often untenable in their implication of a firm distinction in all cases” (Greetham 1992, pp.335-336). Further, there appears to be no widely-applicable typology of edits in digital scholarly editing and collation, with different materials calling for different typologies (see TEI-L 2016).

Our typology of edits departs from previously proposed ones by focusing entirely on types which can be identified automatically, based on surface features. While limited in scope and excluding any semantic criteria, our typology may serve as a first approach to the edits of any text and allow quantitative comparison of some key phenomena. We believe that our method could be applied to other variants of fiction — by itself or incorporated alongside another taxonomy, including accidentals/substantives — particularly to novels which begin as self-published works but are later edited and re-released, an increasingly important phenomenon in contemporary fiction.

Bibliography

Dekker, R. and Middell, G. (2011). Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements*. Supporting Digital Humanities 2011*. University of Copenhagen, Denmark. 17-18 November 2011.

Eve, M. P. (2016). “You have to keep track of your changes”: The Version Variants and Publishing History of David Mitchell's Cloud Atlas, *Open Library of Humanities.* https://olh.openlibhums.org/article/10.16995/olh.82/

Greetham, D. (1992). *Textual scholarship: An introduction.* New York/London: Garland Publishing.

Ho, Y. (2011). *Corpus Stylistics in Principles and Practice: A Stylistic Exploration of John Fowles’ The Magus.* New York: Continuum.

Hunt, J. W. & Mcilroy, M. D. (1975). An algorithm for differential file comparison. *Computer Science.*

Modern Language Association (2011). Reports from the MLA Committee on Scholarly Editions, Guidelines for Editors of Scholarly Editions, available at: https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/Guidelines-for-Editors-of-Scholarly-Editions

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys.* 33 (1): 31–88. doi:10.1145/375360.375365

Schöch, C. (2016). Detecting Transpositions when Comparing Text Versions using CollateX. *The Dragonfly’s Gaze.* http://dragonfly.hypotheses.org/95

TEI-L (2016). Types of Edits. TEI-List. http://tei-l.970651.n3.nabble.com/Types-of-edits-tp4028495.html

van Hulle, D. (2008). *Manuscript Genetics, Joyce’s Know-How, Beckett’s Nohow.* Gainesville: University Press of Florida.

Weir, A. (2011). *The Martian.* Self-published.

Weir, A. (2014). *The Martian.* New York: Crown Publishing Group.

Prototypical Emotion Developments in Literary Genres

Evgeny Kim
evgeny.kim@ims.uni-stuttgart.de
Universität Stuttgart, Germany

Sebastian Padó
sebastian.pado@ims.uni-stuttgart.de
Universität Stuttgart, Germany

Roman Klinger
roman.klinger@ims.uni-stuttgart.de
Universität Stuttgart, Germany

Introduction

Storytelling is a central form of human artistic expression. An ingredient of the appeal of stories is their emotional content. Readers of literature form explicit mental representations of fictional characters’ emotional states (Gernsbacher, Goldsmith, and Robertson, 1992; Vega, 1996). Even more, a gripping “[...] literary work produces a complex emotional experience in the reader. This experience is inseparable from the depictive content of the narrative” (Hogan, 2011). This raises the question of the relationship between the narrative and emotional levels in literature. We explore how computational emotion analysis can contribute to the characterization of story genres, which are difficult to define and for which various criteria have been proposed, including stylistic ones (Biber and Conrad, 2009) and narratological ones (Chatman, 1978).

Our hypothesis is that genres can be linked to the development of predominant emotions over the course of the text. To test this, we present a computational model of multi-label emotion analysis of literary genres and apply it to a set of English literary works from the Project Gutenberg for five genres, namely adventure, romance, mystery, science fiction, and humorous fiction. We identify prototypical shapes for each genre and show that this analysis produces results, which can find a place in the computational analysis of literary genres and extend existing stylometric approaches.

Cuddon (2012) defines adventure as “a form of fiction [...] in which the hero conventionally undergoes a series of testing and episodic adventures” and mystery as a narrative involving the “detection of crime, with the motives, actions, arraignment, judgement and punishment of a criminal”. Baldick (2015) defined Romance as narratives with "improbable adventures of idealized characters”. Today, however, the term covers many forms of fiction, including love stories. We use the term romance as a literary genre in this broader sense. Regarding science fiction stories, it is generally agreed that they are “[...] about an amazing variety of things, topics, ideas. They include trips to other worlds, quests, the exploration of space...” (Cuddon, 2012). Humorous fiction is comical literature “written chiefly to amuse its audience" (Cuddon, 2012).

Methods

We calculate emotion scores for eight basic emotions, namely joy, sadness, trust, disgust, fear, anger, surprise, and anticipation (Plutchik, 2001). We use the NRC Emotion Lexicon (Mohammad and Turney, 2013). Since the data in Project Gutenberg is diachronic, this method of emotion recognition might not be appropriate for older texts and, in general, may suffer from low recall. However, it can be considered a high-precision approach suitable for our purpose.

To obtain an emotion analysis for a story, we start by computing emotion scores for spans of text (Klinger, Suliya, and Reiter, 2016). Formally, we compute the score es(e, S) for an emotion e and a span of tokens S=<tn,...,tm> as

$$es(e,S) = \frac{1}{|D_e|} \sum_{t_i \in S} 1_{t_i \in D_e}$$

where De is a dictionary containing words associated with emotion e and $1_{t \in D}$ is 1 if t ∈ D and 0 otherwise. We do this for each of our eight emotions, obtaining an eight dimensional “emotion vector” for each span. We analyzed the stability of our results across different settings and found that different dictionaries affect the actual values but not the relation between different time steps. These scores are not probabilities, but could be transformed if needed.

To observe development over time, we could use sliding windows; however, continuous time series are notoriously difficult to interpret. Therefore, we adopt a simpler scheme inspired by the five-act theory of dramatic acts (Freytag, 1863), according to which dramas are divided into five acts: exposition, rising action, climax, falling action, and denouement. We consequently divide each text into five successive, equal-sized spans (since different texts have different length, the size of acts varies between texts) that we assume to correspond roughly to dramatic acts in Freytag’s theory, with exposition at position 1 and denouement at position 5, and compute an eight-element emotion vector for each Act.

Experimental Evaluation

We now demonstrate how this emotion aggregation into five acts can contribute to the understanding of different literary genres.

Data

We collect 2113 books from Project Gutenberg that belong to five genres found in the Brown corpus (Francis and Kucera, 1979), namely adventure (585 books), romance (383 books), mystery (380 books), science fiction (562 books), and humorous fiction (203 books). The corpus is available from the authors upon request.

The selection is based on the Library of Congress Subject Headings in the metadata. We select all books that contain the word “Fiction” in combination with one of the following labels: “Adventure stories”, “Love stories”, “Romantic fiction”, “Detective and mystery stories”, “Science fiction” or “Humor”. Furthermore, we reject books with the following labels: “Short stories”, “Complete works”, “Volume”, “Chapter”, “Collection”, “Part”. This constraint is aimed at excluding files which contain partial or multiple works.

Qualitative analysis

Each plot in Figure 1 depicts the act-by-act development for one emotion with their emotion score es(e,S). Since we interpret shapes rather than values, we omit the legend. The average over all books is shown in a dark-colored line. The area around that line corresponds to confidence intervals at a 95% confidence level.

For sadness, anger, fear, and disgust, all five genres show a close correspondence, namely a consistent increase of the emotion from Act 1 through Act 5 – corresponding to gripping narratives. Mystery and science fiction lack the drop in anger and tend to end with higher levels of this emotion. Joy is inverse to these emotions, showing a decreasing tendency from Act 1 to Act 5 for all genres with exception of humorous fiction that shows a plateau between Acts 1 and 4.

In adventures, all emotions increase in the first half of the books, but drop sharply between Act 4 and Act 5. This is consistent with Whetter (2008), according to whom adventures are marked by a late drop in emotional tension when the hero’s misfortunes come to an end. The only exception is trust that shows increase towards the end for all genres, which is especially noticeable in adventures. A potential reason is that prototypical adventure novels are ‘upbeat’ in that they cultivate virtues such as courage and loyalty (Baldick, 2015, p. 5) and depict heroes that do not lose trust even amid unexpected dangers.

The results for anticipation and surprise show less coherent tendencies which we find difficult to interpret. These two emotions appear less constitutive to the narrative structure of genres, at least those that we currently consider: anticipation and surprise can occur under (almost) any circumstances. Mystery fiction has a slightly different pattern, where anticipation exhibits steady increase from Acts 1 to 4 and its peak coincides with the peak for surprise at Act 4.

Quantitative analysis

We analyze the genre-specific time course of emotions quantitatively by computing associations between genres and the Act in which an emotion “peaks” in a story. We define a random variable vie for an emotion peak as vie=1 iff the highest value of emotion e is in Act i. The association between each genre and emotion peaks vie follows point-wise mutual information (Church and Hanks, 1990),

$$PMI(g, v_e^j) = \log \frac{p(v_e^j, g)}{p(g)p(v_e^j)},$$

where probabilities are computed as relative frequencies over the dataset.

Figure 2 gives insight into the genre-specific emotion patterns. For instance, disgust is characteristic of Act 4 for all genres. The only exception is science fiction that does not list disgust or surprise among the top 10 features. Trust is important at the beginning and in the end of adventures and science fiction, but is missing in mystery. Similarly, romance fiction is not characterized by anticipation among top-ranked features, corresponding to its “anticipation” curve that decreases monotone from beginning to end. Interestingly, humor is the only genre that does not contain joy among the top 10 features.

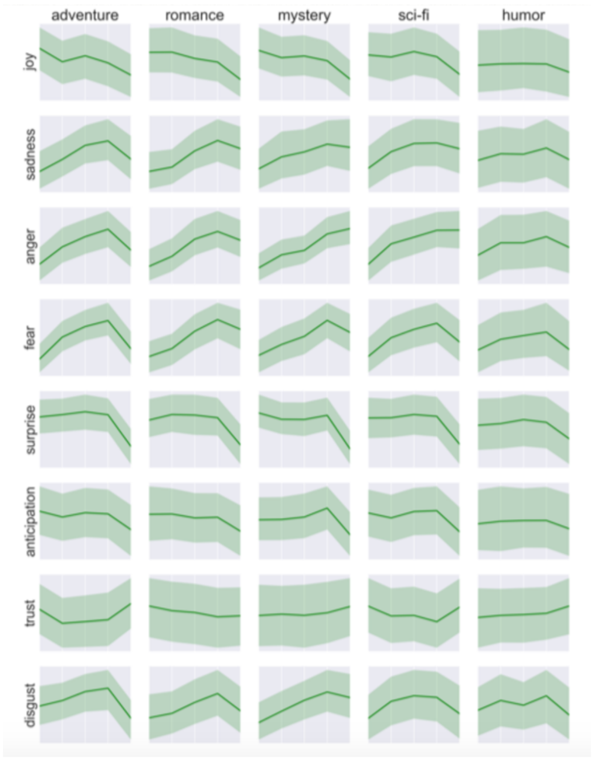


Figure 1: Genre-specific emotion curves

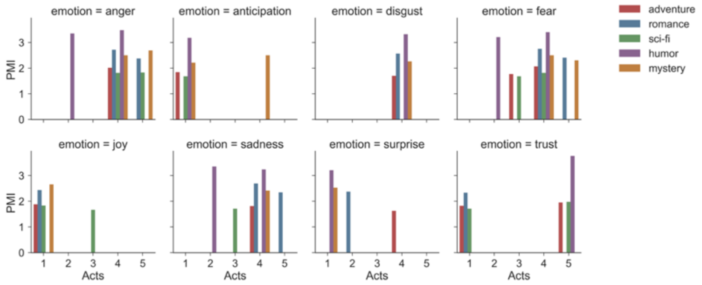


Figure 2: Top 10 PMI features for each genre

Related work

Sentiment and emotion in fiction have been previously addressed computationally by Mohammad (2012), Nalisnick and Baird (2013), Heuser, Moretti, and Steiner (2016), among others. Samothrakis and Fasli (2015) is the only work we are aware of which discusses emotions in context of genres.

The study most related to ours is Reagan et al. (2016). They propose a pipeline that tracks emotions in text. Their main claim is that stories typically follow one out of six “emotional arcs” regarding happiness.

Conclusion, discussion and future work

We investigated the relationship between emotional development in literature and genre and observe differences among emotions. The genre of adventure stands out, especially concerning the end of the story arc. Our results can provide a novel starting point for characterizing similarities and differences within and between literary genres.

Our observations require further investigation regarding the underlying factors. For instance, it might be argued that the pattern for mystery stories is dominated by the subgenre of crime fiction. Future work will combine our distant reading approach with close and scalable reading. Furthermore, to improve emotion recognition, we plan to use distributional methods for expanding the existing lexical resources and adapting them to texts from different historical periods (cf. Buechel, Hellrich, and Hahn, 2016).

Acknowledgements

This research has been conducted within the Center for Reflected Text Analytics (CRETA), which is funded by the German Ministry for Education and Research (BMBF).

Bibliography

Baldick, C. (2015). *The Oxford dictionary of literary terms*. OUP Oxford.

Biber, D. and Conrad S. (2009). *Register, Genre, and Style.* Cambridge University Press.

Buechel, S., Hellrich J., and Hahn U. (2016). Feelings from the Past – Adapting Affective Lexicons for Historical Emotion Analysis. In: *LT4DH 2016,* p. 54.

Chatman, S. (1978). *Story and Discourse: Narrative Structure and Fiction and Film*. Cornell University Press.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), pp.22-29.

***Cuddon, J. A.** (2012). Dictionary of literary terms and literary theory. John Wiley & Sons.*

Francis, W. N. and Kucera H. (1979). *Brown corpus manual. Brown University, Rhodes island.*

Freytag, G. (1863). *Die Technik des Dramas.* Leipzig, Germany: Hirzel.

Gernsbacher, M. A., Goldsmith H. H., and Robertson R. R. (1992). Do readers mentally represent characters' emotional states? *Cognition & Emotion*, 6(2), pp. 89-111.

Heuser, R., Moretti F., and Steiner E. (2016). *The Emotions of London.* Stanford: Stanford Literary Lab, Stanford University. Available at: https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf [accessed 5 Mar. 2017].

Hogan, P. C. (2011). *What literature teaches us about emotion.* Cambridge University Press.

Klinger, R., Suliya S. S., and Reiter N. (2016). Automatic Emotion Detection for Quantitative Literary Studies: A case study based on Franz Kafka's "Das Schloss" und "Amerika. In: *Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Krakow, Poland,* pp. 826-828.

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4), pp. 730-741.

Mohammad, S. M. and Turney P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), pp. 436-465.

Nalisnick, E. T. and Baird H. S. (2013). Extracting sentiment networks from Shakespeare's plays. In: *Document Analysis and Recognition (ICDAR) 2013 12th International Conference on.* IEEE, pp. 758-762.

Plutchik, R. (2001). The Nature of Emotions. Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4), pp. 344-350.

Reagan, A. J., Mitchell L., Kiley D., Danforth C. M., and Dodds P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1), p. 31.

Samothrakis, S. and Fasli M. (2015). Emotional sentence annotation helps predict fiction genre. *PloS One*, 10(11).

Vega, M. (1996). The representation of changing emotions in reading comprehension. *Cognition & Emotion*, 10(3), pp. 303-322.

Whetter, K. S. (2008). *Understanding genre and medieval romance.* Ashgate Publishing, Ltd.

Measuring completeness as metadata quality metric in Europeana

Péter Király
pkiraly@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH
Göttingen
Germany

Introduction

The functionalities of an aggregated metadata collection are dependent on the quality of metadata records. Some examples from Europeana, the European digital library, to display the importance of metadata: (a) Several thousands records have the title „Photo" without further descriptions; how can a user find these objects?, (b) Several data providers listed in the „Institution" facet under multiple different names, should we expect that the user will select all name forms of an organization?, (c) Without formalized date value, we are not able to use the functionality of interactive date range selectors. The question is how can we determine which records should be improved, and which are good enough? The manual evaluation of each record is not affordable. This paper proposes a methodology and a software package, which can be used in Europeana and elsewhere in the domain of cultural heritage.

Background and foundations

Europeana collects and presents cultural heritage metadata records. The database contains more than 53 million records from more than 3200 institutions (figures extracted from the Europeana Search API) in the Europeana Data Model (EDM) schema. The organizations send their data in EDM or in another metadata standard. Due to the variety of original data formats, cataloging rules, languages and vocabularies, there are big differences in the quality of the individual records, which heavily affects the functionalities of Europeana's services.

In 2015 a Europeana task force investigating the problem of metadata quality published a report (Dangerfield et al., 2015), however – as stated – „there was not enough scope … to investigate … metrics for metadata quality ….” In 2016 a wider Data Quality Committee was founded. The current research is conducted in collaboration with it, having the purpose of finding methods, metrics and building an open source tool (see also, the project’s Github page) to measure metadata quality.

State of the art

The computational methods of metadata quality assessment emerged in the last decade in the domain (Bruce and Hillmann, 2004, Stvilia et al., 2007, Ochoa and Duval, 2009, Harper, 2016). Papers defined quality metrics and suggested computational implementations. They however mostly analyzed smaller volumes of records, metadata schemas which are less complex than EDM, and usually applied methods to more homogeneous data sets. The novelty of this research is that it increases the volume of records, introduces data visualizations, and provides open source implementation to use in other collections.

Methodology

For every record, features were extracted or deducted which somehow related to the quality of the records. The main feature groups are:

- **simple completeness** – ratio of filled fields,
- **completeness of sub-dimensions** – fields groups support particular functions, such as searching, or accessibility,
- **existence and cardinality of fields** – which fields are filled and how intensively.

The measurements happen on three levels: on individual records, on subsets (e.g. records of a data provider), and on the whole dataset. On second and third level we calculate aggregated metrics; the completeness of structural entities (such as the main descriptive part and the contextual entities – agent, concept, place, timespan – connecting the description to linked open data vocabularies).

The final completeness score is the combination of two approaches. In the first one the weighting reflects sub-dimensions. In the second one, the main factor is the normalized version of cardinality to prevent biasing effect of extreme values.

The tool – built on big data analytics software Apache Spark, the R statistical software and has a web front-end – is modular. There is a schema-independent core library and schema specific extensions. It is designed to be used in continuous integration for metadata quality assessment.

Results

Comparison of the scores of the field importance and field cardinality approaches shows that they give different results (however they correlate by the Pearson’s coefficient of 0.52.). Because of the nature of calculation the compound score is quite close to the first approach: the functionality based scores lie in the range of 0.186 and 0.76 and cardinality scores are in the range of 0.031 and 0.335, and it has smaller effect on the final score.

There are data providers, where all (in some cases more than ten thousand) records get the same scores: they have uniform structure. The field-level analysis shows (what one simple score is not able to testify) that in these collections all the records has the very same (Dublin Core based) field set. On the other end there are collections where both scores diverge a lot. For example in the identifying sub-dimension a data provider has five distinct values (from 0.4 to 0.8) almost evenly distributed while one of the best collection (of the category) is almost homogeneous: 99,7% or the records have the same value: 0.9 (even the rest 0.3% has 0.8). It means that in the records of the first dataset the corresponding fields (dc:title, dcterms:alternative, dc:description, dc:type, dc:identifier, dc:date, dcterms:created and dcterms:issued in the ore:Proxy part and edm:provider and edm:dataProvider in the ore:Aggregation) are frequently not available, while they are almost always there in the second. The tool provides different graphs and tables to visualize the distribution of the scores.

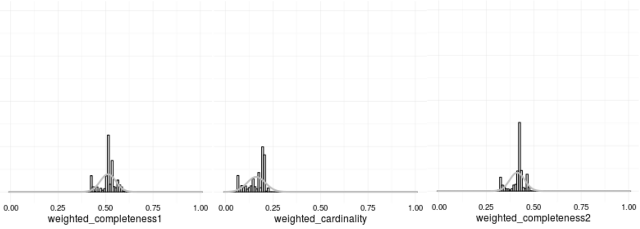


Figure 1. Distribution of completeness scores in a dataset. We can see the differences between the functionality based (left),the cardinality based (center) and the combined method (right).

From the distribution of the fields the first conclusion is that lots of records miss contextual entities, and only a couple of data provider has 100% coverage (6% of the records has *agent*, 28% has *place*, 32% has *timespan* and 40% has *concept* entities). Only the mandatory technical elements appear in every records. There are fields, which are defined in the schema, but not filled in the records and there are overused fields – e. g. *dc:description* is frequently used instead of more specific fields (such as table of contents, subject related fields or alternative title).

Users can check all the features on top, collection, and records level on the web interface. Data providers get a clear view of their data, and based on this analysis they can design a data cleaning or data improvement plan.

Europeana is working on its new ingestion system which integrates the tool. When a new record-set will arrive, the measuring will run automatically, and the Ingestion Officer can check the quality report.

Further work

We will examine other metrics (e.g. multilinguality, accuracy, information content, timeliness), and check known metadata anti-patterns. We plan to compare the scores with experts' evaluation and with usage data and to implement related W3C standards: Shapes Constraint Language (Knublauch and Kontokostas, 2016), and Data Quality Vocabulary (Albertoni and Isaac, 2016).

Conclusion

In the research we re-thought the relationship between functionality and the metadata schema, implemented a framework which proved to be successful in measuring structural features which correlate with metadata issues, and we were able to select low and high quality records. We remarkably extended the volume of the analyzed records by introducing big data tools, which were not mentioned previously in the literature.

I showed my research in case of a particular dataset and data schema but the method I follow based on generalized algorithms, so it is applicable to other data schema. Several DH researches based on schema defined cultural databases, and in those cases the research process could be improving by finding the weak points of the sources.

Acknowledgements

I would like to thank all of the members of the Europeana Data Quality Committee.

Bibliography

Dangerfield, M-C. et al. (2015). "Report and Recommendations from the Task Force on Metadata Quality." (http://pro.europeana.eu/files/Europeana_Professional/Publications/Metadata%20Quality%20Report.pdf)

Bruce, T. R. and Hillmann, D. I. (2004). "The Continuum of Metadata Quality: Defining, Expressing, Exploiting." In Hillman, D. and Westbrooks E. (eds), *Metadata in Practice*, Chicago, ALA Editions, 2004.

Stvilia, B., Gasser, L., Twidale, M. B. and Smith, L. C. (2007). "A framework for information quality assessment." *Journal of the American Society for Information Science and Technology*, 58(12): 1720-1733.

Ochoa, X. and Duval, E. (2009). "Automatic evaluation of metadata quality in digital repositories." *International Journal of Digital Libraries*, 10: 67-91.

Harper, C. (2016). "Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA)." *The Code4Lib Journal*, 33 (http://journal.code4lib.org/articles/11752)

Knublauch, H. and Kontokostas, D. (eds.) (2016). "Shapes Constraint Language (SHACL). W3C Working Draft 14 August 2016." (https://www.w3.org/TR/shacl/)

Albertoni, R. and Isaac, A. (eds.) (2016). "Data on the Web Best Practices: Data Quality Vocabulary. W3C Working Group Note 30 August 2016" (https://www.w3.org/TR/vocab-dqv/)

A Pale Reflection of the Violent Truth? Practice and Pedagogy with a Digital Geography of American Lynching

Seth Kotch
sethk@unc.edu
UNC Chapel Hill, United States of America

Elijah Gaddis
ejgaddis@live.unc.edu
UNC Chapel Hill, United States of America

We propose a discussion of the pedagogical process and representational implications of our digital project, *The Red Record*. *The Red Record* is an open-ended, ongoing project created using DH Press that situates historic American lynchings, starting with those that took place in North Carolina, on a Cartesian map. In its current state the project offers primary source documentation and scholar-produced context for each of the 112 lynching events, which took the lives of 147 people in the state between 1865 and 1941. The initial project was undertaken by a seminar of first year students at the University of North Carolina, and subsequently amended with the research of two additional undergraduate classes. In the session proposed here, we detail the process of research and the intertwined workflow that simultaneously teaches students primary research skills and forces them to confront the legacies of racial violence inherent in the landscapes they live in. We suggest that the pedagogical model of this project has broader implications for the ways that digital representations might seek to engage diverse populations in understanding and contributing to the knowledge of the places they are from. In the papers here, we consider both the implications of digital representations of lynching, and the possibilities and practical obstacles in extending this project through student and community crowdsourcing.

Until recently, North Carolina has enjoyed a reputation as a less objectionable southern state, whether as a "vale of humility" between the less commendable Virginia and South Carolina, or as a "business progressive" southern state, concerned enough about outside investment to control its citizens' worst impulses. The problem of lynching, widely studied in states such as Georgia and Virginia, has received less attention from North Carolina's scholars: the only monograph on the subject is a summary-study our research found to contain a number of errors. Visual scholarship on the subject is limited. In 2000, *Without Sanctuary*, a gruesome collection of lynching postcards, opened in New York and was soon followed by a website which still exists but which shows its age. *Without Sanctuary* was and is an important project, one that shocked viewers as to the extent of historical violence. Yet today, as we regularly witness violence against persons of color on social media or television news, we must wonder about troubling side effects of *Without Sanctuary*: that in presenting lynchings as sepia-toned horrors of the past, it appeared to relegate such violence to history and allow for a self-satisfied reassurance that this time had passed.

Such is the challenge for any project addressing historical trauma. Adding to this concern are questions of representation. In uncritically presenting images of lynchings, *Without Sanctuary* risked reproducing the work of those postcards, which traveled the country presenting damaged black bodies that titillated as much as it horrified, and even in their extremity regularized and commodified anti-Black violence. We know that lynchings themselves echoed across generations in Black communities (Wolf, 1992) did *Without Sanctuary* address that trauma or amplify it? Do the victims of lynchings have, in a sense, the right to be forgotten? (Causey, 2015) In pursuing Locating Lynching, we remained aware of these risks and encountered still more: how to choose visual anchors for our display, whether or not to publish images of trauma, how to share accounts of lynchings from the white and Black presses without false equivalence, what kinds of information to include about persons who do not deserve to be remembered only as victims. Our panel will address these issues in detail, share our solutions, and explain our next steps as we continue to work.

A key element of our project and our panel is pedagogy, both in terms of engaging undergraduate students with digital, primary-source research and in terms of engaging them with a traumatic subject. It is the rare undergraduate student, particularly in a seminar designed for first-year students, who is well-equipped to conduct original research. But with the recent digitization of hundreds of thousands of pages of small-town newspapers, of censuses from the 19th and early 20th centuries, and of city directories and historic maps, the hidden histories of lynchings, their victims, and their perpetrators have never been more visible. As historians, we can and have learned stories from the darkened margins: We have learned the story a young man who escaped a lynching attempt and fled the state, ending up, married and with a son, as a doorman in the United Nations building in Manhattan. We have learned the names of the white men who kidnapped from jail an African American man jailed on a rape accusation and hanged him. We know these people's addresses, their occupations, the names of their parents and children. In both these marginal stories, and in the aggregating potential of digital visualizations, we can tell some small part of the abbreviated life stories of these people, albeit in the context of their deaths. Teaching undergraduate students the process whereby they can make these discoveries – as detailed in our nine-page research guide – can be transformative. We ask our students, who are awash in digital representation and misrepresentation, to carefully consider their choices in saving, commenting on, and presenting past identities.

Bibliography

Causey, M. (2015) "The Right to Be Forgotten and the Image-Crimes of Digital Culture" in M. Causey et al., *Performing Subject in the Space of Technology* (New York: Palgrave,), 69-81.

Wolf, C. (1992) "Constructions of a Lynching," *Sociological Inquiry*, Vol. 62, No. 1, 83-97.

A Survey on Research Data at the Faculty of Arts and Humanities of the University of Cologne

Simone Kronenwett
simone.kronenwett@uni-koeln.de
Cologne Center for eHumanities, Germany

Brigitte Mathiak
mathiak@gmail.com
Cologne Center for eHumanities, Germany

Executive summary

In the course of the digitization of information, research data management has become one of the most important new areas of research. Universities have to prepare themselves to provide their academics and researchers with the necessary infrastructures and services. To identify the current demands regarding the handling of research data at the Faculty of Arts and Humanities of the University of Cologne, the Data Center for the Humanities conducted an online survey in 2016 in cooperation with the Office of the Dean of the Faculty of Arts and Humanities as well as the University and Library of Cologne. The enquiry aimed to characterize the present situation and to obtain information on the demands in the sectors research data management and consultation services. Our talk will show ongoing developments at the international and national level in research data management, present the results of the survey and discuss potential conclusions.

Relevance

One of the most important fields of action in research which is developing along with the digitalization of information is research data management (RDM). The universities face the challenge of offering their researchers adequate structures and services. The managing board of the

German universities organized in the German Rector's Conference (HRK) has identified this as a key task (HRK 2014 and 2015). Moreover, in the recently published position paper called "Performance by diversity" the German Council for Scientific Information Infrastructures (RfII) makes a series of recommendations concerning how research data should be managed in the future (RfII 2016). The RfII was tasked by Germany's Joint Science Conference (GWK) with formulating broad-based recommendations for the science system in Germany as a whole. In addition, according to estimates by the German Research Foundation (DFG), up to 90% of the digital generated research data and results are still getting lost (Winkler-Nees 2011, p. 5) or "disappear in the drawer" (Kramer 2014) shortly after completion of research projects and are therefore not available for further use and reuse.

Method

As a structured way to gain information, we decided to follow the six stages process recommended for survey research (Müller et al. 2014). In addition, the survey is based on the relevant articles published in the handbook "Methods of Library and Information Science" by Umlauf, Fühles-Ubach & Seadle (Umlauf et al. 2013). The six steps are briefly explained below.

The Internet survey, the online questionnaire as well as the detailed report are published on the DCH-Website (see also Kronenwett 2017).

Research goals and constructs

The goal of the survey is to contribute to the conceptual development of the Data Center for the Humanities (DCH), which was founded as a central infrastructure service institution by the faculty dealing with humanities research data in 2013. In practice, the enquiry aims to characterize the present situation and to obtain information on the demands in the sectors RDM and consultation services offered by the DCH in cooperation with the University and City Library of Cologne (USB), one of the local partners of the DCH. Another goal was the comparability with other surveys conducted in the field.

Population and sampling

Because RDM should be handled in a way specific to each discipline (Sahle et al. 2013), the survey targeted only researchers at the Faculty of Arts and Humanities of the University of Cologne - one of the largest humanities faculties in Europe. The survey's population is limited to the academic staff of the Faculty of Arts and Humanities of the University of Cologne. In particular, the survey focused on researchers who are responsible for data-driven research projects.

Questionnaire design and biases

Firstly, with regards to content, conceptual and methodical design of the survey, Internet surveys and online questionnaires on research data at national and international scientific institutions and research institutes were analyzed so far available (the website "forschungsdaten.org" offers an overview regarding national and international surveys on research data). Secondly, the questionnaire design was tailored and adapted to suit the unique circumstances which can be found at the Faculty of Arts and Humanities (DCH 2016, CCeH 2016). Finally, the results of a series of expert interviews carried out by the DCH with researchers at the Faculty were taken into account (Blumtritt 2016, p. 16). The questionnaire addresses five issues, namely (1) research data, (2) use of data archives, (3) support for research data, (4) discipline and position, (5) interest.

Review and survey pretesting

After a review of the questionnaire with stakeholders, such as the dean’s office, the library and the data protection officer, a test link was sent to 20 potential subjects. This included representatives of all subject groups of the Faculty of Arts and Humanities (Faculty of Arts and Humanities 2016) as well as external experts (mostly sociologists and colleagues with RDM-background). After several feedback loops, the questionnaire was further modified and optimized.

Implementation and launch

The questionnaire was compiled using Kronenwett & Adolphs online survey tool (Kronenwett & Adolphs 2017). It was put online from 2016-05-30 to 2016-06-12 (2 weeks). Depending on individual answers the questionnaire contained up to 24 questions.

Data analysis and reporting

The questionnaire was completely answered by 136 participants (out of 191 persons who started the survey) which is 71.20% completion rate. The following selection of data analysis and reporting takes into account only these participants (n=136).

Results

Our objective in the compilation of the questionnaire was to answer the following questions:

1. What research data are available?
2. What is the need for research data?
3. What support do the members of the Faculty of Arts and Humanities want from the DCH?

Regarding the first question, sustainability and data volume were important to us. As far as sustainability is concerned, the majority of respondents are storing their research data on their local computers: 70% work computer, 70% private computer, multiple responses were possible (see fig. 1). Only 14% are storing their data in a data archive, a number that is also reflected in other questions like how many participants can imagine their data being stored in a data archive.

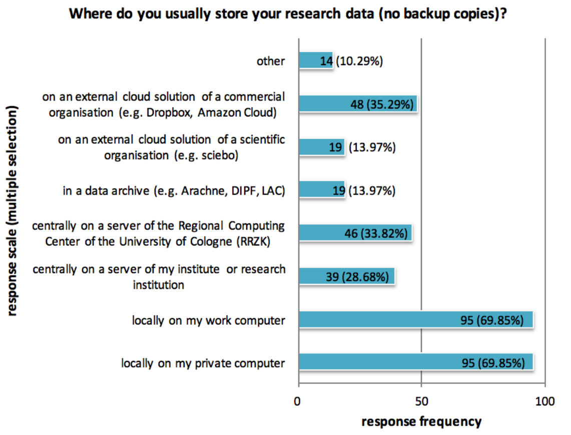


Figure 1. Storage places (n=136)

Regarding sustainability standards the given answers are fatal results since structured access and retrievability of research data are only ensured in professional data archives. Cloud solutions are also quite popular (35% use by commercial vendors and 14% of scientific vendors) because they ensure overall data access and data share. But regarding aspects like data plausibility, traceability or even long-term preservation they are totally unsuitable.

This result could be explained by the fact that the participants do not reflect their approach to sustainability and traceability. The vast majority of the respondent's self-assessment regarding their own skills in RDM is rated to be average or even less (71%) (see figure 2).

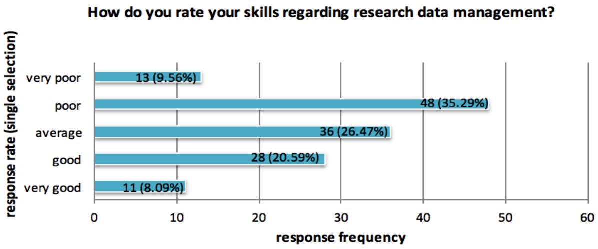


Figure 2. Self-assessment RDM-skills (n=136)

Sustainability is seen as a problem. 66% of the participants state that data could be lost when there is no one to be responsible for the website. 60% fear problems with data conversion. There is some sensibility towards the issues of finding the data (45%) and documentation of the data (41%). Figure 3 shows all answers concerning problems the participants see with preserving research data. Interesting is that both privacy and data theft are the concerns voiced the least frequent (11% each), despite the fact that we encounter these concerns frequently in our consulting practice. But this could be an anomaly due to us asking from the user perspective rather than the data giving perspective which is more typical to our consulting.

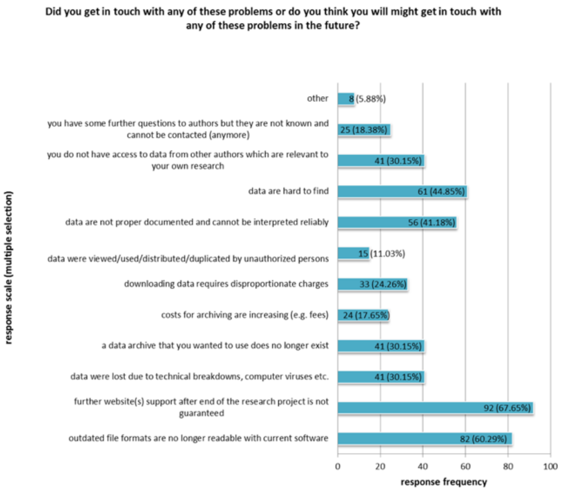


Figure 3. Past and future problems (n=136)

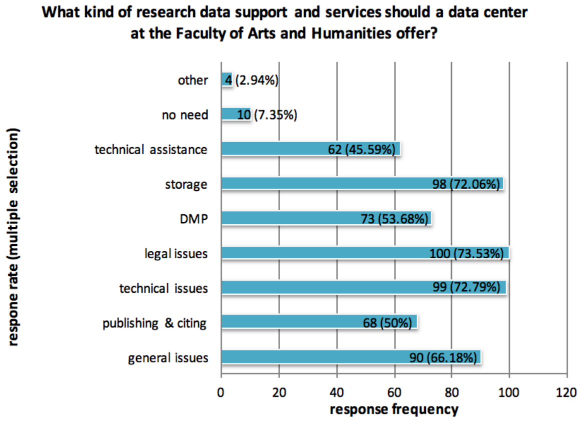


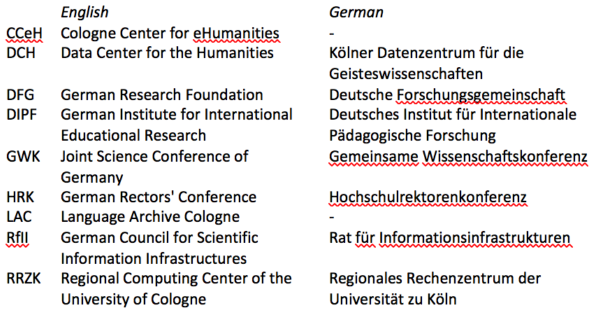
Figure 4. Support and services needed (n=136)

In an effort to improve our services towards the faculty, we also asked which services should be provided by the DCH. Here legal and technical issues featured prominently (74% and 73% respectively; also cf. Fig. 4). Requests for storage (72%) and consultation (66%) on general issues are also in high demand.

Conclusion and outlook

As a result of the survey, we propose the following recommendations for action for the University of Cologne on the one hand and for the DCH on the other hand. Together with the local library (USB), the DCH now offers legal counseling on the subject of research data with a specialized lawyer. We are currently planning a project for improving the sustainability of living software systems, since the survey showed that this is an eminent problem in our faculty. Based on the projected storage space from the survey, we have negotiated with the computing center to provide that space centrally for all the members of our faculty.

In our talk, we will give more details on the study and its results and will also compare it to other surveys conducted internationally. We feel that surveys of this nature are an important tool to shape strategic decisions made in institutions concerned with research data.



	English	German
CCeH	Cologne Center for eHumanities	-
DCH	Data Center for the Humanities	Kölner Datenzentrum für die Geisteswissenschaften
DFG	German Research Foundation	Deutsche Forschungsgemeinschaft
DIPF	German Institute for International Educational Research	Deutsches Institut für Internationale Pädagogische Forschung
GWK	Joint Science Conference of Germany	Gemeinsame Wissenschaftskonferenz
HRK	German Rectors' Conference	Hochschulrektorenkonferenz
LAC	Language Archive Cologne	-
RfII	German Council for Scientific Information Infrastructures	Rat für Informationsinfrastrukturen
RRZK	Regional Computing Center of the University of Cologne	Regionales Rechenzentrum der Universität zu Köln

Figure 5. List of abbreviations used.

Bibliography

Kronenwett & Adolphs (n.d.), http://www.kronenwett-adolphs.com/en (accessed on 2017-03-26)

Averkamp, S., Gu, X., Rogers, B. (2014). Data management at the University of Iowa: A university libraries report on campus research data needs. University of Iowa, http://ir.uiowa.edu/lib_pubs/153/ (accessed on 2016-10-31)

Bauer, B., Ferus, A., Gorraiz, J., Gumpenberger, C., Gründhammer, V., Maly, N., Mühlegger, J. M., Preza, J. L., Sánchez Solís, B., Schmidt, N., Steineder, C. (2015). Researchers and their data. Results of an Austrian survey - report 2015. e-infrastructures austria, Vienna, http://e-infrastructures.at/das-projekt/deliverables (accessed on 2017-03-12)

Blumtritt, J. (2016). Consulting Workflow for Humanities Research Data, Talk at Humanities Research Data Conference (FORGE 2016): Beyond Data - Sustainability for Research Application and Software, 2016-09-15, University of Hamburg, https://www.gwiss.uni-hamburg.de/gwin/ueber-uns/forge2016/praesentationen/f1606-blumtritt-mathiak.pdf (accessed on 2017-03-12)

Cologne Center for eHumanities (Ed.) (2016). Digital Humanities. Structures - Teaching- Research, Cologne, http://cceh.uni-koeln.de/broschure-digital-humanties-2016/ (accessed on 2016-10-12)

Data Center for the Humanities (n.d.) http://dch.phil-fak.uni-koeln.de (accessed on 2017-03-26)

Data Center for the Humanities (2016). Research Data Survey 2016, http://dch.phil-fak.uni-koeln.de/umfrage-2016.html (accessed on 2017-03-26)

Faculty of Arts and Humanities (Ed.) (2016). Research 2016/17, Cologne

Faculty of Arts and Humanities, University of Cologne (n.d.) Subject Groups Overview, http://phil-fak.uni-koeln.de/9785.html (accessed on 2017-03-26)

Forschungsdaten.org, (n.d.) Overview of Surveys on Research Data at Scientific Institutions, http://www.forschungsdaten.org/index.php/Umfragen_zum_Umgang_mit_Forschungsdaten_an_wissenschaftlichen_Institutionen (accessed on 2017-03-26)

Hochschulrektorenkonferenz (2014). Management von Forschungsdaten als strategische Aufgabe der Hochschulleitungen. Empfehlung der 16. HRK-Mitgliederversammlung am 13. Mai 2014 in Frankfurt am Main, https://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_01.pdf (accessed on 2017-03-12)

Hochschulrektorenkonferenz (2015). Wie Hochschulleitungen die Entwicklung des Forschungsdatenmanagements steuern können. Orientierungspfade, Handlungsoptionen, Szenarien. Empfehlung der 19. Mitgliederversammlung der HRK am 10. November 2015 in Kiel, https://www.hrk.de/uploads/tx_szconvention/Empfehlung_Forschungsdatenmanagement__final_Stand_11.11.2015.pdf (accessed on 2016-08-12)

Kramer, B. (2014). Datenflut an Unis. Forscher müssen teilen lernen, in: Spiegel Online, 26.02.2014, http://www.spiegel.de/unispiegel/jobundberuf/umgang-mit-daten-der-glaeserne-forscher-a-954958.html (accessed on 2016-10-31)

Kronenwett, S. (2017). Forschungsdaten an der Philosophischen Fakultät der Universität zu Köln (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft, Bd. 78) Köln

Kuipers, T., van der Hoeven, J. (2009). Insight into digital preservation of research output in Europe. Survey report. PARSE insight, European Commission, Brussels, http://parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf (accessed on 2017-03-12)

Müller, H., Sedley, A., Ferrall-Nunge, E. (2014). Survey research in HCI, in: J. S. Olson, W. A. Kellogg (Ed.): Ways of Knowing in HCI, New York, p. 229-266

Rat für Informationsinfrastrukturen (2016). Leistung aus Vielfalt [Performance by diversity]. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, http://www.rfii.de/?wpdmdl=1998 (accessed on 2017-03-12)

Ray, J. M. (Ed.) (2014). Research Data Management: Practical Strategies for Information Professionals, West Lafayette/USA

Sahle, P., Kronenwett, S. (2013). Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center for the Humanities', in: LIBREAS. Library Ideas, 23, p. 76-96, http://edoc.hu-berlin.de/libreas/23/sahle-patrick-1/PDF/sahle.pdf (accessed on 2017-03-12)

Schöpfel, J. Prost, H. (2016). Research data management in social sciences and humanities: A survey at the University of Lille (France), in: LIBREAS. Library Ideas, 29, p. 98-112, http://nbn-resolving.de/urn:nbn:de:kobv:11-100238193 (accessed on 2017-03-12)

Winkler-Nees, S. (2011). Vorwort, in: S. Büttner, H.-C. Hobohm, L. Müller (Ed.): Handbuch Forschungsdatenmanagement, Bad Honnef, p. 5-6, http://www.forschungsdatenmanagement.de (accessed on 2016-08-30)

Huma-Num : Une infrastructure française pour les Sciences Humaines et Sociales. Stratégie, organisation et fonctionnement

Nicolas Larrousse
nicolas.larrousse@huma-num.fr
Centre National de la Recherche Scientifique, France

Olivier Baude
olivier.baude@huma-num.fr
Centre National de la Recherche Scientifique, France

Adeline Joffres
adeline.joffres@huma-num.fr
Centre National de la Recherche Scientifique, France

Stéphane Pouyllau
stephane.pouyllau@huma-num.fr
Centre National de la Recherche Scientifique, France

Introduction

La recherche en sciences humaines et sociales vit un tournant numérique qui provoque une évolution sensible des pratiques de recherche.

Aujourd'hui, les chercheurs produisent d'importants volumes de données numériques et utilisent des outils spécialisés pour y accéder, les manipuler, les visualiser et les diffuser. Cela nécessite d'associer ces grands volumes de données à des moyens technologiques qui se doivent d'être stables et conséquents. La maitrise de ce nouvel environnement doit s'appuyer sur le développement de compétences diversifiées et en constante évolution.

La mise en œuvre de cet écosystème ne peut plus être réalisée avec les moyens dont disposent les chercheurs individuellement ou même à l'échelle d'une équipe de recherche. Une infrastructure, au sens élargi du terme, est nécessaire pour être à même de fournir et mutualiser différents services et d'accompagner les équipes de recherche tout au long de leurs projets numériques.

La TGIR Huma-Num est une infrastructure de recherche française au service de la recherche en sciences humaines et sociales. Elle est bâtie sur une structure originale destinée à proposer des services dédiés à la production et à la réutilisation de données numériques mais aussi à favoriser et accompagner l'émergence d'outils et de bonnes pratiques issus des communautés scientifiques des sciences humaines et sociales. Enfin, Huma-Num est étroitement associée à la construction d'infrastructures européennes permettant tout autant l'échange et la valorisation de services que le renforcement de réseaux d'expertises.

Notre présentation s'articule autour des trois grands axes développés par Huma-Num pour répondre aux besoins de la recherche en sciences humaines et sociales.

- La création et le soutien de consortiums, cercles d'expertise disciplinaires ou non, permettant une concertation collective par les communautés ;
- Le développement de services technologiques destinés à l'outillage des données de la recherche;
- L'ouverture à l'international par la participation à des projets d'infrastructures Européennes mais aussi par des échanges bilatéraux plus ciblés.

Les consortiums réunissent des acteurs issus de structures et de projets de recherche divers, autour de thématiques, d'objets ou de méthodes communs. L'objectif est de favoriser l'appropriation des dispositifs numériques par ces communautés et ainsi de créer des synergies en lien avec les services de Huma-Num. Le spectre des activités de ces consortiums est large, allant, par exemple, de la création d'applications jusqu'à la définition de schéma de métadonnées, en passant par la production de guides de bonnes pratiques associées à des formations spécifiques. Actuellement, Huma-Num soutient les activités de huit consortiums, principalement disciplinaires, qui s'articulent avec le réseau des Maisons des Sciences de l'Homme (réseau géographiquement réparti sur les campus universitaires et du CNRS en France métropolitaine) et dont certains collaborent à des projets européens.

Les services gérés directement par l'équipe de la TGIR Huma-Num sont destinés à fournir des outils tout au long du cycle de vie des données : de la production de données brutes jusqu'à leur préservation après traitement. Pour pouvoir proposer des ressources bien dimensionnées et adaptées, Huma-Num n'a pas créé son propre centre de données, mais s'appuie sur les compétences et les moyens de centres de calcul nationaux français.

L'un des objectifs principaux est de mettre en œuvre un cadre d'outils permettant aux chercheurs d'exploiter au mieux leurs données mais surtout de les amener à les partager et, in fine, à les préserver. Il s'agit de permettre l'appropriation par les chercheurs de méthodes et services numériques au cœur des processus de recherche scientifique. Les technologies d'interopérabilité, en particulier celles proposées par le Web Sémantique, sont au cœur de cette chaine de traitement permettant des échanges fluides entre les différentes briques qui la composent. Le Web Sémantique permet aussi de décloisonner les données de la recherche par l'utilisation de référentiels et normes communs.

Ces outils appuient une démarche générale d'accompagnement des utilisateurs. Ainsi, lors d'une demande d'utilisation, Huma-Num effectue de manière systématique une sensibilisation à la pérennisation, à la curation, au partage des données et aux enjeux de la normalisation.

Ce point est crucial dans le fonctionnement de cette grille de services qui doit aussi permettre la montée en compétence des communautés.

Les activités internationales de Huma-Num sont de différente nature mais elles visent essentiellement à valoriser les dispositifs français d'infrastructure et de la recherche française et à construire un dialogue scientifique et des dispositifs dépassant les frontières nationales. Huma-Num porte ainsi la participation française à des ERICs (European Research Infrastructure Consortium), disciplinaires comme CLARIN depuis 2017, ou à vocation plus large pour les Sciences Humaines en incluant le domaine culturel comme DARIAH depuis sa création en 2014. Ces ERICs, établis sur une longue durée, ne sont pas financés directement par la Commission Européenne mais par les États participants. L'originalité du budget de l'ERIC DARIAH est que 90% de celui-ci est constituée de contributions « en nature » : l'objectif étant de ne pas dupliquer à l'échelle européenne des services qui existeraient déjà au niveau national. Le rôle d'Huma-Num est donc d'identifier les services et expertises nationaux, en particulier ceux en provenance des consortiums et des grands opérateurs français de l'information scientifique et technique et du patrimoine culturel, et d'en faire un ensemble cohérent capable d'intégrer et de compléter l'offre européenne constituée par les partenaires. En parallèle, Huma-Num participe à des projets à plus court terme destinés à soutenir la construction de ces grandes infrastructures et qui sont, eux, financés par la Commission Européenne (dans le cadre des programmes H2020).

Enfin, hors Europe, Huma-Num maintient des liens étroits avec les mondes francophones (e.g. Québec) et hispanophones (e.g. Amérique du Sud). Là aussi, l'objectif est de valoriser les services d'Huma-Num ainsi que ceux issus des communautés nationales et de bénéficier en retour d'échanges d'expertises, notamment dans les domaines de la préservation et de la curation de données issues de la recherche en Sciences Humaines et Sociales.

La TGIR Huma-Num a élaboré sa définition de services technologiques autour du cycle de vie des données. Cette présentation vise à présenter, à partir d'une démarche réflexive, les services d'Huma-Num et la pertinence de leur articulation avec les autres composantes de l'infrastructure que sont les consortiums ainsi que leur projection au niveau international.
En effet, Huma-Num propose des outils adaptés à chaque étape du cycle des données de la recherche :

- Du simple stockage de sauvegarde au début du projet ; Suivant le taux prévisible d'utilisation des données, il est proposé plusieurs solutions technologiques : les données « froides » (i.e. peu utilisées) sont gérées à moindre coûts sur des stockages distribués alors que les données « chaudes » sont mises à disposition via des stockages performants de type NAS ;
- Des outils de traitement des données. Ceux-ci sont mutualisés comme par exemple des logiciels de SIG (Système d'Information Géographique) qui seraient trop couteux à acheter pour une utilisation ponctuelle par un projet. Pour ces opérations de traitements, la TGIR met à disposition une grande puissance de calcul ;
- De l'hébergement Web pour présenter les données et permettre leur partage et leur accès ainsi qu'un travail collaboratif. Pour des besoins plus spécifiques des Machines Virtuelles sont mises à disposition afin d'offrir une grande souplesse d'exploitation ;
- Enfin, et c'est là l'originalité de l'infrastructure, il est proposé un ensemble de services qui permet la diffusion, la citation, la pérennisation et surtout la promotion de la réutilisation de ces données:
    - NAKALA pour stocker des données documentées et les partager, associé à NAKALONA pour les éditorialiser
    - ISIDORE pour diffuser ces métadonnées en les enrichissant, les classifiant et les positionnant dans le LOD (Linked Open Data)

Pour les données qui le nécessitent (e.g. des données de type patrimonial), un service de pérennisation à long terme (plus de 20 ans) en partenariat avec le CINES dont c'est la mission, complète l'offre.

La pertinence des demandes d'ouverture de services est évaluée par un comité interne à Huma-Num qui s'appuie également sur les recommandations de son conseil scientifique.

Cet ensemble de services peut être visualisé de manière synthétique à cette adresse:
http://www.huma-num.fr/services-et-outils

Par ses différentes composantes, la TGIR Huma-Num s'emploie à développer des réponses aux besoins nouveaux des différents acteurs de la recherche provoqués par l'utilisation du numérique. Le fil rouge qui les relie est de pouvoir rendre possibles de nouvelles recherches dans le domaine des sciences humaines et sociales et, au-delà, de constituer de manière dynamique un savoir partagé.

Mining the Cultural Memory of Irish Industrial Schools Using Word Embedding and Text Classification

Susan Leavy
susan.leavy@ucd.ie
University College Dublin, Ireland

Emilie Pine
emilie.pine@ucd.ie
University College Dublin, Ireland

Mark T. Keane
mark.keane@ucd.ie
University College Dublin, Ireland

Introduction

The Industrial Memories project aims for new distant (i.e., text analytic) and close readings (i.e., witnessing) of the 2009 Ryan Report, the report of the Irish Government's investigation into abuse at Irish Industrial Schools. The project has digitised the Report and used techniques such as word embedding and automated text classification using machine learning to re-present the Report's key findings in novel ways that better convey its contents. The Ryan Report exposes the horrific details of systematic abuse of children in Irish industrial schools between 1920 and 1990. It contains 2,600 pages with over 500,000 words detailing evidence from the 9-year-long investigation. However, the Report's narrative form and its sheer length effectively make many of it findings quite opaque. The Industrial Memories project uses text analytics to examine the language of the Report, to identify recurring patterns and extract key findings. The project re-presents the Report via an exploratory web-based interface that supports further analysis of the text. The methodology outlined is scalable and suggests new approaches to such voluminous state documents.

Method

A web-based exploratory interface was designed to enable searching and analysis of the contents of the Report represented within a relational database. The relational structure detailed the categories of knowledge contained in the Report along with key information extracted from the text (Figure 1). The Ryan Report is composed of paragraphs containing an average of 87 words. These paragraphs were represented as database instances and annotations detailing semantic content were linked through the relational structure. Named entities were automatically extracted using NLTK (Looper and Bird, 2002).

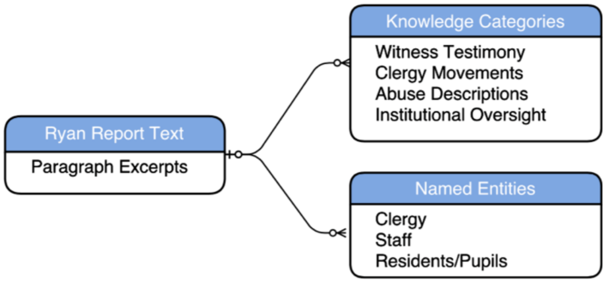


Figure 1: Knowledge Database Relational Structure

Classifying Paragraphs into Different Knowledge Categories

The Ryan Report describes key elements of an enduring system of abuse that operated in Irish industrial schools. Its paragraphs tend to focus on particular topics, allowing them to be classified and annotated. For instance, some cover the extent and nature of abuse, others present witness testimony, report on institutional oversight or on how clergy were moved from one school to another in response to allegations. By classifying paragraphs in terms of these high-level knowledge categories it becomes easier to put a shape on many of the report's findings and to analyse it to provide new readings.

Some of these paragraph-categories were identified using automated text classification. Others were extracted using a rule-based search (e.g., excerpts on institutional oversight). In building classification models, a variety feature sets were examined using a random forest classifier along with manually selected test data. A bag-of-words approach to feature selection yielded results that were over-fitted due to the small samples of training data. However, feature selection based on context-specific semantic lexicons generated from a sample of seed-words using a word embedding algorithm was found to yield accurate results. Lexicons were generated using the word2vec algorithm developed by Mikolov (2013) following an approach identifying synonyms outlined by Chanen (2016).

Movements of Staff and Clergy (Transfer Paragraphs)

An important paragraph-category covers those dealing with the Catholic Church's response to allegations of abuse. The typical response to discovered abuse was to transfer clergy from one institution to another, only for the abuse to re-occur (e.g., *"...Br Adrien was removed from Artane and transferred to another institution..." (CICA Vol. 1, Chapter 7, Paragraph 829))*. Such transfers are described in many different ways in language that often obscures what was happening (e.g., transfers out of the Order, effectively sackings, are described as "dispensations to be released from vows"). We carried out a "by-hand" analysis to find transfer-paragraphs using verb-searches and then expanded this set using machine-learning classifiers.

Initial readings of the Report suggested a set of verbs frequently used to describe the transfer of staff and clergy, including 'transfer', 'dismiss', and 'sack'. The highest-ranking similar words reoccurring over five word2vec models were then identified. Features based on this lexicon, along with names of schools and clergy were extracted from 250 training examples (Table 1). A classification model then classified unseen text from the Report.

Text Category	**Features Extracted**
Direct Speech	Reporting verbs, personal pronouns, punctuation (colons, quotation marks, commas, question marks, contractions), newlines
Movements of Staff and Clergy	Transfer verbs, names of clergy, schools
Descriptions of Abusive Events	Clergy and staff, parts of body, abusive action, emotions and implements associated with abuse

Table 1: Optimal Features Extracted from Report Identifying Witness Testimony

Witness Testimony (Witnessing Paragraphs)

Witness testimonies in the Ryan Report are indicated through reporting verbs and structural speech markers (e.g., punctuation). Using these features, Schlör et. al. (2016) gained accuracy of 84.1 percent in automatically classifying direct speech. Reporting verbs in the Ryan Report are often specific to its context such as apology, allegation or concession. To extract these from the text, highest-ranking similar words across multiple word embedding models were identified based on seed terms generated from WordNet, 'said', 'told' and 'explained'. The resulting context-specific synonyms combined with WordNet synonyms formed a lexicon of reporting verbs tailored to the language of the Report (Table 2). A classification model was developed using these features along with punctuation information using 500 training examples.

Seed Words	Context Specific Lexicon for Reported Speech				
said	answered	alleged	warned	enounced	explained
told	learned	recounted	claimed	verbalise	believed
explained	confirmed	surmised	denied	verbalised	added
	described	relieved	asserted	assure	replied
	say	protested	witnessed	articulate	thought
	tell	stating	called	apologise	knew
	told	describes	informed	pardon	felt
	state	agreed	said	pardoned	recalled
	posit	admitted	explained	remember	saying
	posited	convinced	advised	articulated	told
	submit	presumed	assured	enounce	thinks
	submitted	screams	tells	condoned	remembered
	express	reported	requested	condone	conceded
	expressed	complained	heard	saw	realised
	narrate	asking	says	explicate	stated
	narrated	commented	confessed	explicated	insisted
	recount	questioned	remarked	apology	guarantee
	recite	accepted	alleged	concluded	asked
	recited	recollection	suggested	mentioned	

Table 2: Context Specific Synonyms Using Word Embedding

Descriptions of Abusive Events (Abuse Paragraphs)

To evaluate the scale of abuse throughout the industrial school system, excerpts from the Report detailing abusive events were extracted. The language describing abuse incorporates a broad range of linguistic features. A set of seed-words from which to base a semantic lexicon for feature extraction, was not immediately apparent on reading the Report. A support vector machine algorithm was therefore used to extract the most discriminative features based on a sample set of 200 paragraphs.

Analysis of the support vectors showed that terms distinguishing excerpts describing abuse formed five categories: abusive actions, body parts, emotions engendered in the victims, implements and names of staff and clergy. Sample words associated with each category were then used as seed-words to generate word embedding models to extract similar terms from the Report. Features based on these five lexicons, combined with names of clergy and staff were then used to generate a predictive model of abusive events.

Findings and Conclusions

This research demonstrates how word embedding can be used to compile context-specific semantic lexicons to extract features for text classification. These features allowed paragraphs for each knowledge category to be automatically classified based on manually selected training data.

Classification	No. Classified Excerpts
Clergy Movements	1,340
Direct Speech	1,920
Abusive Events	1,365

Table 3: Total Number of Classified Paragraphs

The performance of classifiers was evaluated using 10-fold cross-validation on the training data and showed high levels of accuracy in categorisations (Table 4).

Classification	Precision	Recall	F-Score	Accuracy
Clergy Movements	.91	.91	.91	91.2%
Direct Speech	.94	.93	.93	93.6%
Abusive Events	.93	.93	.93	93.3%

Table 4: Performance on Training Data: Random Forest Classifier. Weighted Average Results Using 10-fold Cross-Validation

The classification models were applied to unseen data and performance evaluated by manually inspecting classifications of 600 randomly selected excerpts from the Report as shown in Table 5. Though overall accuracy levels remained high, precision of the classifications did fall somewhat, especially in relation to identifying speech and transfers.

Classification	Precision	Recall	F-Score	Accuracy
Clergy Movements	.58	1.0	.73	92%
Direct Speech	.84	.92	.88	94%
Abusive Events	.86	.88	.87	95%

Table 5: Performance on Unseen Text: Classification of Report Evaluated on Random Samples

Error analysis showed that incorrectly classified excerpts (false positives and negatives) were commonly those where the meaning of the language was subtle or vague. Paragraphs incorrectly classified as quoted speech for instance, were in fact quotations from letters and diary entries. Unidentified speech excerpts all consisted of short quoted phrases.

Transfers of clergy were reliably detected. However, there was a high rate of false positives due to the fact that the transfer of children throughout the school system is described using similar language (e.g. "*The witness remembered … when he was leaving Artane at nine years of age…" (CICA Vol. 1, Ch. 7, Paragraph 466)*). Classifying excerpts describing abuse yielded few false positives but it also returned the highest levels of false negatives. In these instances, references to abuse was subtle or addressed emotional abuse. As such, it was necessary to manually filter results.

This paper has demonstrated that machine learning can be used to classify text based on a limited number of examples, when used in conjunction with word embedding to generate context-specific semantic lexicons. Re-presenting the Ryan Report in the form of a relational database with a web-based exploratory interface has facilitated comprehensive analysis of the Report, and has exposed new insights about the dynamics of the system of child abuse in Irish industrial schools. In reformulating how the Ryan Report can be presented, this research presents a scalable approach to digital analysis of state reports.

Acknowledgements

This research is part of the Industrial Memories project funded by the Irish Research Council under New Horizons 2015.

Bibliography

Chanen, A. (2016). Deep learning for extracting word-level meaning from safety report narratives. In *Integrated Communications Navigation and Surveillance (ICNS), 2016* (pp. 5D2-1). IEEE.

Loper, E. and Bird, S. (2002). NLTK: e Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1.* (ETMTNLP '02). Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 63–70.

Mikolov, T., et al. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26 (NIPS2013),* pp. 3111–119.

Schöch, C., Schlör, D., Popp, S., Brunner, A., Henny, U., & Tello, J. C. (2016). Straight Talk! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. In *Digital Humanities 2016* (pp. 346-353).

Real and Imagined Geography at City-Scale: Sentiment Analysis of Chicago's "One Book" Program

Ana Lucic
alucic@depaul.edu
DePaul University, United States of America

John Shanahan
jshanah1@depaul.edu
DePaul University, United States of America

Since fall 2001, the Chicago Public Library (CPL) has chosen fiction and nonfiction around which to organize city-wide public events, book discussions, and other creative programming. This "One Book One Chicago" (OBOC) program has been a successful ongoing civic initiative with great public visibility, with participants ranging from the Mayor of Chicago to countless book group volunteers across the city. Our "Reading Chicago Reading" project—supported by the National Endowment for the Humanities Office of Digital Humanities and Microsoft—works to discover how text characteristics, library branch demographics, and promotional activities are linked variables that can be used to predict patron response to future OBOC titles. The OBOC program acts as a recurring experiment in data capture, for each chosen work represents a probe into library usage and, by extension, a window onto the elective reading behavior of the diverse patrons of a major American city.

This paper will report comparative circulation data for three recent OBOC choices that are Chicago-centered and three that are not Chicago-centered. The three Chicago-centered books are:

1) *The Adventures of Augie March* (1953) by Saul Bellow
2) *The Warmth of the Other Suns* (2011) by Isabel Wilkerson
3) *The Third Coast* (2014), by Thomas Dyja

The three recent non-Chicago OBOC choices are:

1) *The Book Thief* (2007) by Markus Zusak
2) *The Amazing Adventures of Kavalier and Clay* (2000) by Michael Chabon
3) *Gold Boy, Emerald Girl* (2010) by Yiyun Li

We are keen to answer whether a Chicago setting and, more particularly, particular measures of linguistic sentiment about Chicago people and places, have measurable influence on the popularity of books across Chicago. Specifically, we are interested in examining the question of whether "Chicago" books, fictional and nonfictional, checked out in greater numbers when they feature characters, events, and places situated close to the readers' own neighborhood library branch? (We use CPL library branch as a proxy for patron home address, which we cannot know from the library system's anonymized checkout data.)

The results of this analysis have the potential to provide empirical answers to long-standing questions in digital humanities research: in his Atlas of the European Novel 1800-1900, for example, Franco Moretti speculated that perhaps "fictional spaces are particularly suited to happy endings," but did not have hard numbers to judge one way or the other at the time (18 n.6). More recently, in The Bestseller Code, Jodie Archer and Matthew Jockers argue that "while it does matter whether an author chooses a city or the wilderness, the specific city does not matter all that much when it comes to bestselling" (227). Our sentiment analysis findings will contribute to open research questions: maybe in fact the city matters when readers are in that city, and when the places and people in that same city are written about in particular linguistic registers. If literary form and real geography do have detectable ties to one another, our project ought to be able to capture the effect.

To compare the circulation pattern of non-Chicago and Chicago-related OBOC books, we used one year of city-wide circulation data for each book, starting with the date of the title's public announcement as the OBOC choice. This data was normalized by dividing the circulation raw numbers with the total number of visitors for that year and multiplying the result by 1000. Normalizing by the number of circulated copies was sometimes difficult because some branches did not have any copies (but could borrow them from other branches). Given that libraries oftentimes allocate books based on the size of the library and based on the number of visitors, we decided to normalize by the overall number of visitors. Distribution patterns for Chicago and non-Chicago related sets of books are represented through the histograms and QQ plots in Figure 1. As this analysis indicates, the circulation distribution across 79 Chicago library branches does not follow bell-shape distribution and is positively skewed. The Wilcoxon signed rank test was used to test to what degree the difference in the distribution for these two sets can be attributed to chance.

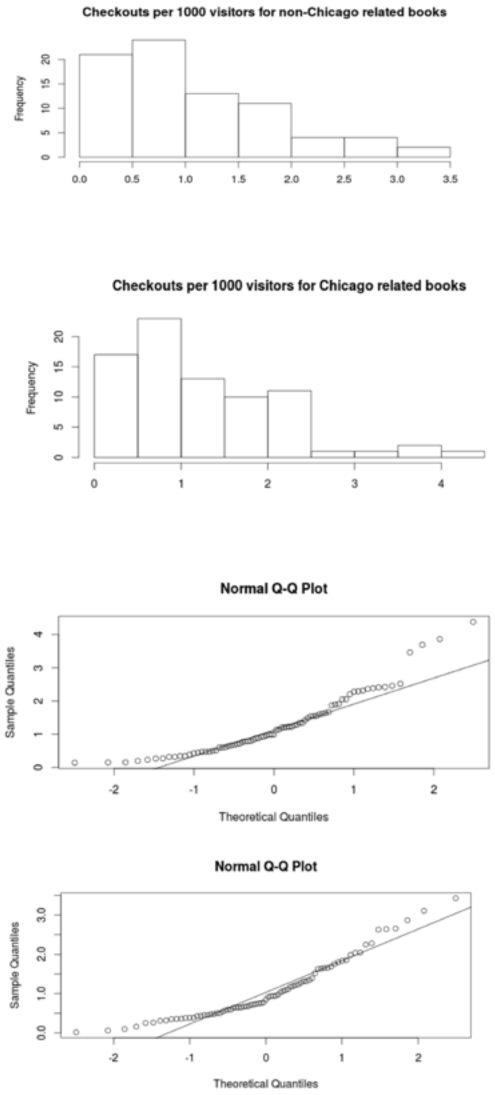


Figure 1. On top, histograms representing checkouts per 1000 visitors for non-Chicago related and Chicago related books. Below, the QQ plots for non-Chicago and Chicago related books checkouts.

Results indicate that the probability that the difference in the circulation distribution across 79 Chicago library branches for the two (paired) sets of books can be attributed to chance is very low (p < .01).

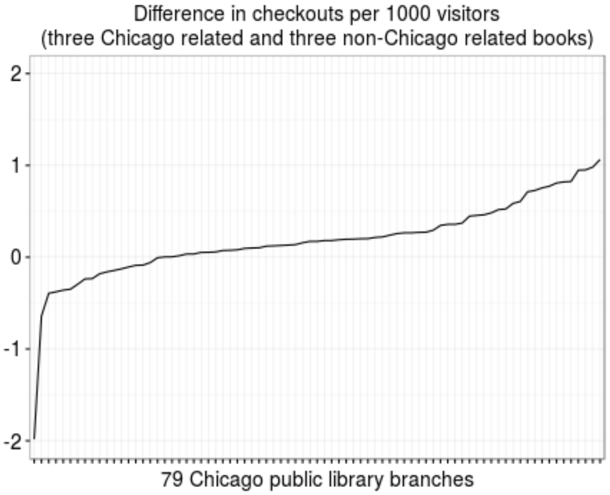


Figure 2. The y-axis indicates the difference in the checkouts (1-year of circulation data) for three Chicago related and three non-Chicago related books.

The y-axis in Figure 2 represents the difference between the checkouts per 1000 visitors for Chicago related and non-Chicago related books. Figure 2 indicates that the three non-Chicago related books circulated more than the set of Chicago related books in some library branches in the Chicago area (where the line drops into negative difference). In some branches, however, the difference is almost minimal. The plot also indicates that, in some branches, the OBOC Chicago-related choices had, in fact, more checkouts than the non-Chicago OBOC choices (where the difference is positive). Although it is difficult to establish which factors contribute to this difference in circulation and although we cannot attribute this difference between the two distributions to the mere fact that one set contains references to Chicago whereas the other does not, we plan to represent the library branches that are associated with a larger number of checkouts for Chicago non-related books and those that are associated with a larger number of related books on the Chicago map and analyze them against the sociodemographic and socioeconomic characteristics of different branches (obtained from the American Community Survey data). In the future, we plan to add more Chicago related books to the analysis and observe how this may affect this observed pattern.

A further question of interest to us is, do the sentiment measures for these texts map in consistent ways for different neighborhoods? To examine these questions, we rely on Stanford CoreNLP natural language processing capabilities (Manning et al., 2014). Given that the identification of places and locations is important for our analysis, we use a tool that has consistently achieved good rankings and, in general, boasts superior accuracy rates when compared to other named entity recognizers (Rodriquez et al., 2012; Atdağ & Labatut, 2013): the Stanford Named Entity Recognizer, a part of the CoreNLP suite of tools. Before running the named entity recognizer, the text is first tokenized into sentences using the NLTK sentence tokenizer. The CoreNLP program then tokenizes sentences into words, identifies lemma for each individual word, uses the Penn Treebank part of speech information (Toutanova & Manning, 2000), and also notes Persons, Locations, Time Reference, and Numbers in the sentences (Finkel et al., 2005). We are specifically interested in locations as this category would not only identify Chicago as a location but also its streets and landmark buildings. For sentiment analysis, we are using the Stanford sentiment analysis tool (Socher et al., 2013)—also part of the Stanford CoreNLP—to annotate each sentence with the sentiment score on the following scale: Very Positive, Positive Neutral, Negative, Very Negative.

Preliminary analysis on the sentiment associated with sentences that contain the word Chicago in the three Chicago-related books is indicated in Figure 3:

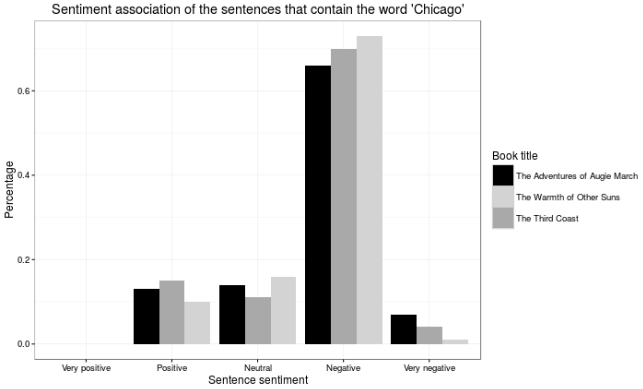


Figure 3 Sentiment distribution of sentences that contain the word Chicago in Augie March, The Warmth of Other Suns, and The Third Coast

The raw count of sentences with their sentiment ratings was normalized by the total number of sentences that contain the word Chicago. Noticeable in Figure 3 is that although these three books differ according to genre, and although they differ in terms of topical coverage and date of publication, we see a rather similar sentiment score pattern with respect to sentences that contain the word Chicago. We suspect that this overall similarity pattern will start to change as we dig deeper into the location data: we must note here that this initial analysis above does not yet take into account local references such as Pizzeria Uno, Pullman, the South Side, Monroe Street, and the like, but do not also use the word Chicago in the sentence. We plan to obtain a set of place names associated with Chicago through resources such as Open Street Maps and GeoNames and search for all the occurrences of Chicago place names. Additionally, we plan to use the indexes in the back of some of the books as trusted sources of local place names.

Bibliography

Archer, J. and Jockers, M. (2016). *The Bestseller Code: Anatomy of the Blockbuster Nove*l. New York: St. Martin's Press.

Atdağ, S., & Labatut, V. (2013). A Comparison of Named Entity Recognition Tools Applied to Biographical Texts. In *2nd international conference on systems and computer science* (pp. 228–233). https://arxiv.org/ftp/arxiv/papers/1308/1308.0661.pdf.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistic*s.

Moretti, F. (1998). *Atlas of the European Novel 1800-1900*. London: Verso.

Rodriquez, K. J., Bryant, M., Blanke, T., & Luszczynska, M. (2012). Comparison of named entity recognition tools for raw OCR text. In *KONVENS* (pp. 410-414). http://www.oegai.at/konvens2012/proceedings/60_rodriquez12w/60_rodriquez12w.pdf

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Paper presented at the EMNLP.

Toutanova, K., and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Paper presented at the Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, Hong Kong.

Modeling Interpretation in 3DH: New dimensions of visualization

Jan Christoph Meister
jan-c-meister@uni-hamburg.de
University of Hamburg, Germany

Johanna Drucker
drucker@gseis.ucla.edu
UC Los Angeles, United States of America

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
University of Alberta, Canada

Introduction

Visualization techniques developed in the sciences normally focus on the (re)presentation of empirical data. But how can we graphically express *interpretations*? This paper presents the intellectual framework underpinning the 3DH project (Three-dimensional Visualizations for the Digital Humanities), a collaborative project conducted at the University of Hamburg from 2016 to 2019. The project foregrounds data interpretation and develops a visualization paradigm from the epistemological perspective of the humanities. The "third dimension" required in DH visualization techniques is therefore not merely that of an additional quantitative z-axis. Rather, it is an axis that can 'unflatten' (Sousanis 2015) the objectivist notion of visualized data. In our presentation, we will do three things:

- **Digital and visual turn:** Review existing visualization paradigms that emphasize the representational approach. We start with the epistemological issues raised by the digital and visual turn.
- **Visual modelling:** Outline and discuss an interpretative modelling alternative through two case studies of existing tools, CATMA and Voyant, and Temporal Modelling, a platform for creating data through graphical means.
- **"Hermeneuticizing" visualization:** Discuss the design of a full visual framework. We will present possible conventions and prototypes that use them. These inform our case studies and the envisaged infrastructure.

Case studies in our presentation will be drawn from CATMA (a collaborative mark-up & text analysis environment), Voyant (a text analysis platform), and humanities research projects using base images (historical maps) and original models (for non-standard chronologies).

The digital and the visual turn: a hermeneutic ceterum censeo

For centuries, academic discourse in humanities disciplines has relied predominantly on text. In DH, however, visualizations increasingly claim the status of arguments and proofs that play a decisive role in the development and presentation of ideas, findings, and conclusions.

The visual and the digital turn have thus gone hand in hand – but the way in which this synergy manifests itself remains constrained in a symptomatic way. We can print a chart or render it on screen just as we can print or display a text in various media, but we normally cannot subject the chart to in-depth critique in the way we can question and respond to the text. Inadvertently, once generated and communicated as 'output', visualizations seem to take on a quasi-dogmatic quality – they are hard to deconstruct, let alone reconfigure; they state their case but seem removed from critical reflection.

Most current DH visualizations are thus epistemological one-way avenues toward knowledge, from data via rendering algorithm to visual display. Charts, graphs, interactive maps, timelines, and similar representations are by and large imports from the natural and social sciences (Friendly 2008). Many of them emanate from domains of empirical research that conceptualize knowledge production as a function of empirical observation and objective measurement followed by analysis, inference, and conclusion. These approaches to visualization, however, hide two critical aspects, namely

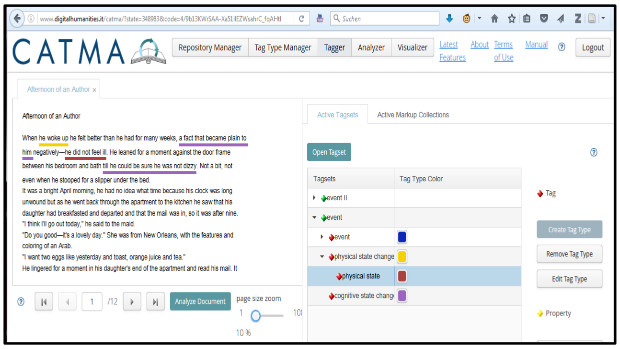
(a) the underlying human modeling of the represented phenomena *as* data, which is already an interpretive and meaning-*creating* act that often oscillates repeatedly between observation and interpretation (Kitchin 2014), and
(b) the meaning-*lessness* of certain visual effects that are owed to contingent technological constraints (screen size, rendering, scaling, choice of color, etc.).

DH is in a unique position to investigate the domains of human experience and of its expression in symbolic practices and artefacts from two complementary methodological vantage points: the numeric, which models them as statistical phenomena, and the hermeneutic, which explores them as phenomena of *meaning* and thus by definition as a function of interpretation (Rockwell & Sinclair 2016). Where *meaning* comes into focus, our theories, object models, and practices must therefore be conceptually aligned and 'hermeneuticized' – just as numeric approaches come with the pre-requisite of quantification. Against this backdrop, we propose to reintroduce the dimension of interpretation into visualization: Methodological principles of hermeneutic approaches, such as multi-perspectivity, subjectivity, and context-boundedness present a challenge which representational visualization cannot and which interpretational visualization must meet.

Two questions arise: What are the defining principles of a genuinely humanistic and hermeneutically oriented approach to visualization? And how can we graphically express and support interpretation in DH visualizations – both as an *activity* and as a *product* of humanistic enquiry?

Visual modeling of interpretation vs. visualization of data

In the 3DH project, we address the former question by conceptual analysis and critique of existing approaches to visualization in DH, and then by systematically specifying and developing a visualization environment that can support higher level data interpretation rather than base-level data representation. In the presentation, we will share our survey of existing tools and their affordances but focus on two tools that we have developed, CATMA and Voyant.



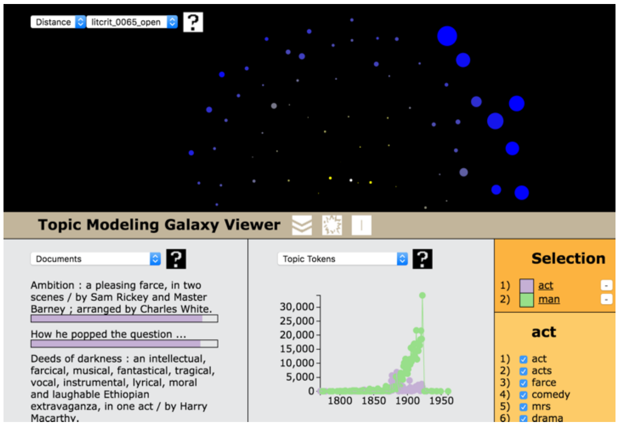
Afternoon of an Author

When he woke up he felt better than he had for many weeks, a fact that became plain to him negatively—he did not feel ill. He leaned for a moment against the door frame between his bedroom and bath till he could be sure he was not dizzy. Not a bit, not even when he stooped for a slipper under the bed.
It was a bright April morning, he had no idea what time because his clock was long unwound but as he went back through the apartment to the kitchen he saw that his daughter had breakfasted and departed and that the mail was in, so it was after nine.
"I think I'll go out today," he said to the maid.
"Do you good—it's a lovely day." She was from New Orleans, with the features and coloring of an Arab.
"I want two eggs like yesterday and toast, orange juice and tea."
He lingered for a moment in his daughter's end of the apartment and read his mail. It

Tagsets	Tag Type Color
event II	
event	
event	🟦
physical state change	🟨
physical state	🟥
cognitive state chang	🟪

Figure 1: Visualization of interpretive text annotation in CATMA

Our premise is that interpretation happens through the deliberate activity of an individual engaging with an image, text, display, or other artifact *to create an argument about its meaning and a way it should be read*. For example, in CATMA (Figure 1) such an activity – in this instance the interpretive act of text annotation – is executed and represented by (a) highlighting a string on screen, (b) assigning it a tag, and (c) storing the annotation in a stand-off markup file. However, the annotation is at the same time (d) visually expressed as colored underlining. Moreover, via its visual representation on screen – the colored underlining – the markup data can also be (e) inspected, analyzed, manipulated directly, and even (f) enriched with meta-annotation. This is but one example of interpretative modeling.



Ambition : a pleasing farce, in two scenes / by Sam Rickey and Master Barney ; arranged by Charles White.

How he popped the question ...

Deeds of darkness : an intellectual, farcical, musical, fantastical, tragical, vocal, instrumental, lyrical, moral and laughable Ethiopian extravaganza, in one act / by Harry Macarthy.

Figure 2: Galaxy Viewer

Current representational 'one-way' techniques like topic modeling (see Figure 2) are seen as a way to deal with scale, they process large amounts of data into summary abstractions called topics that can be displayed as lists or in other ways (Montague et. al 2015). In our second case study, we will therefore show how we are adapting scale tools to create a prototypical bi-directional 3DH visual modeling environment for big data. We believe visual modeling can support not only interpretative close reading of primary data but also the reading of large collections like the collections of the Hathi Trust.

'Hermeneuticizing' base-level visualization through activators: the 3DH framework of interpretive parameters and dimensions

A key goal of the 3DH project is to develop a set of generic graphic features that can be used to create interpretative attributes and/or inflections of visual representations of data, alter underlying data structures, and activate three-dimensional space in the service of interpretative activity. These features which aim to 'hermeneuticize' visualizations are termed *activators*. In the presentation we will show the framework of the activator set that was developed during a series of *charettes* (design workshops) in 2016.

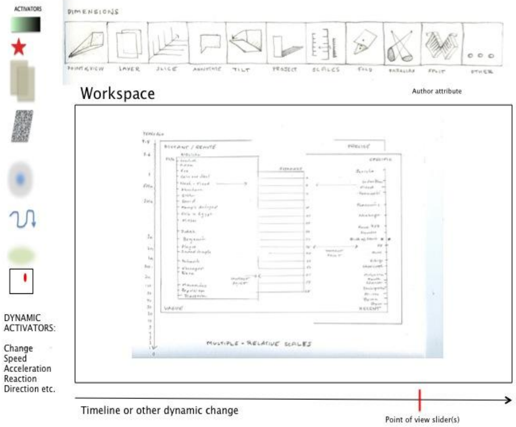


Figure 3: Framework of Concept Modeling workspace: Shows features, activators, and dimensions from various pictorial conventions.

The visual activators in our feature set are not simply graphical marks or animations on a screen display: They perform data structuring functions and as such provide a conceptual framework for 'hermeneuticizing' existing base-level data visualization techniques (see Fig.3). The individual features of this framework indicate and facilitate interpretative moves made by the user, such as a qualification of visualized data structures in terms of salience, irrelevance, uncertainty, degree of completeness, and other attributes or inflections. For example, uncertainty can be expressed by overlaying a standard graph with visual effects such as blur or shading, whereas the introduction of additional interpretative dimensions, such as point of view systems, parallax, relative scales, and other conventions from the visual arts, will support higher levels of interpretative critique and reflection, such as explicitly marking the historicity and context-dependency of underlying data.

Conclusion

As Pinker (1990) argues, the ease with which a particular graph can be understood is a function of the processing effort that goes into the exercise: The more we can rely on 'hard-wired' encoding connections between the visual and the conceptual and the more we are guided by established graph and comprehension schemata (such as Gestalt phenomena), the less 'intelligent' effort we have to put into reading a graph. Yet in a humanities perspective such conventionalized 'ease of comprehension' is a double-edged sword: It may optimize the process of (re)cognition – but it also progressively obscures the constructedness of a visualization, turning it into an apparently self-evident object of perception. The 3DH project counters this anti-hermeneutic tendency toward reification by moving from a conceptualization of the principles of visualization as *interpretative* modeling to the development of a visual language framework, and finally the instantiation of the principles and language in two case studies. In terms of implementation, this approach is supported by drawing on Bertin's *Semiology of Graphics* and the high-level object-oriented *Grammar of Graphics* approach outlined by Wilkinson (2005), and features from game engines, three-dimensional modelling, and other pictorial conventions (Panofsky (1991) and Burgin (1991)).

To conclude, we will discuss next steps toward developing a 3DH environment that can act as a generic, project independent infrastructure for introducing user parameterized enunciative functionality into graphical displays. This infrastructure will make it possible to inscribe into visualizations the critical features of authorship, speaking/spoken subject, and an epistemological perspective grounded in situated and constructed approaches to knowledge. These interpretative principles are well mapped in, e.g., critical theory, narratology, visual studies, and cultural studies, but they have not been integrated into a graphical environment for hermeneutic practice yet: the methodological lacuna which the 3DH project tries to address.

Bibliography

Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, WI, University of Wisconsin Press.

Burgin, V. (1991). "Geometry and Abjection". In: J. Donald (ed.), *Psychoanalysis and Cultural Theory: Thresholds.* London, Macmillan Education, pp. 11–26 .

Chandrasekaran, B. & Lele, O. (2010). "Mapping Descriptive Models of Graph Comprehension into Requirements for a Computational Architecture: Need for Supporting Imagery Operations". In: A. K. Goel, M. Jamnik & N. H. Narayanan (eds.), *Diagrammatic Representation and Inference. 6th International Conference, Diagrams 2010, Portland, OR, USA, August 9–11, 2010. Proceedings.* Berlin & Heidelberg, Springer Verlag, pp. 235–242.

Drucker, J. (2011). "Humanities approaches to Graphical Display". In: *DHQ*, *Digital Humanities Quarterly,* 5 (1). http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html [March 17 2017].

Drucker, J. (2014). *Graphesis,* Cambridge, Harvard University Press.

Friendly, M. (2008). "A Brief History of Data Visualization". In: C.-H. Chen, W. K. Härdle & A. Unwin (eds.), *Handbook of Data Visualization*. Heidelberg, Springer-Verlag, pp. 1–34.

Kath, R., Schaal, G. S. & Dumm, S. (2016). „New Visual Hermeneutics". In: *Cybernetics and Human Knowing*, 23 (2), pp. 51–75.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Los Angeles, SAGE.

Montague, J., Simpson, J., Brown, S., Rockwell, G. & Ruecker, S. (2015). "Exploring Large Datasets with Topic Model Visualization". Conference paper at DH 2015, University of Western Sydney, Australia.

Panofsky, E. (1991). *Perspective as Symbolic Form;* C. Wood, trans.; New York, Zone Books.

Pinker, S. (1990). "A Theory of Graph Comprehension". In: R. Feedle (ed.), *Artificial Intelligence and the future of testing*. Marwah, NJ, Erlbaum Hillsdale, pp. 73–126

Rockwell, G. & Sinclair, S. (2016). *Hermeneutica*. Cambridge, MS & London, MIT Press.

Sousanis, N. (2015). *Unflattening*. Cambridge, MS & London, Harvard University Press.

Wilkinson, L. (2005). *The Grammar of Graphics*. 2nd ed.; New York, Springer.

Regrowing Egyptian cults: The potential of using modern computational methods in the study of ancient religions

Adam Mertel
mertel.adam@gmail.com
Masaryk University, Czech Republic

Tomáš Glomb
tomas.glomb@gmail.com
Masaryk University, Czech Republic

Zdeněk Stachoň
zstachon@gmail.com
Masaryk University, Czech Republic

Introduction

Historical sources provide us only a fragmentary testimony of historical processes. The survival of specific sources and the information they contain is very often determined by chance. Because of these circumstances, the reconstruction of many historical processes remains problematic or, in some cases, almost impossible. In addition to these problems, some long-term historical processes are very difficult to recognize from the perspective of "event history" due to their gradual nature. The interdisciplinary project GEHIR (Generative Historiography of Religion) at Masaryk University strives to, within the framework of the historically oriented study of religions, adopt innovative methods used in the study of the dynamics of complex systems, i.e. methods including mathematical and geographical modelling, agent-based simulations or network analysis. Within the study of historical processes these formalized methods are conceptualized as an innovative third way through which the limitations of the traditional inductive analysis of historical sources and deductive application of social-scientific and cognitive theories to social and historical phenomena can be overcome. In this paper we would like to introduce the results from a case study within the GEHIR which focused on uncovering the possible factors influencing the early spread of the cults of Isis and Sarapis in the ancient Mediterranean.

Methods

In this project, traditional methods of historical research such as critical analysis of the archeological and literary sources or categorization of the historical evidence in the form of relational database were supplemented by the following computational methods:

- **Network analysis** is currently used in various areas of historical research, mainly to understand and analyse complex structures, e. g. trading cities in state, social groups. There are a number of historical research papers that have used network analysis; the work of Rivers et al (2012) should be mentioned.
- **Environmental/geographical modelling** is used to reconstruct the potential of locality within an examined phenomenon that is even partially spatially-dependent. It offers a unique opportunity for researchers to understand how spatial conditions are related to other events. The work of Turchin et al (2013) and others could be used as an illustration of such methods.
- **Mathematical statistical analysis** aims to find hidden logical patterns and existing correlations within multiple phenomena that may subsequently be interpreted in the context of the research questions. Various methods and approaches can be used, including clustering, multidimensional analyses, and exploratory data analysis.

Case study: Egyptian cults in the Aegean Sea region

Early in the Ptolemaic era, in the period between ca. 305 - 167 BCE, the Egyptian cults, particularly those of Isis and Sarapis, spread successfully to ports in the ancient Mediterranean. The reasons behind this process are, however, only partially understood. The original and still respected hypotheses in the academic discussion emphasize either the maritime trade (Fraser, 1960) or Ptolemaic political propaganda (Cumont, 1911:78-80) as key factors in the spread of these cults. Both of these claims are supported by historical evidence. Ptolemaic Egypt was one of the main exporters of grain, Isis was a patron goddess of sailors and many cities in the ancient Mediterranean had close diplomatic relations with the Ptolemies.

To specify our area of interest, this research is mainly focused on the early spread of the Egyptian cults in the area of the Aegean Sea, particularly the Aegean islands. There are several reasons for this selection:

- the main trading routes between Alexandria and continental Greece passed through that particular area with the Aegean islands as potential places of Egyptian interest.
- the first Ptolemies were politically invested in the Aegean using the islands as strategic locations for military bases and administered the Island league.
- islands are more isolated worlds and thus more suitable for modelling.

In order to clarify these hypotheses, we sequentially applied the described computational methods to the issues of maritime transport network reconstruction, approximation of the urgency of importing goods, evaluation of positional advantages of particular islands, and statistical correlation of different modelled factors.

The first task was to reconstruct the maritime transport network, as this was the backbone of ancient Mediterranean trade. For this purpose, Pascal Arnaud's collection of ancient maritime routes were scanned and geo-referenced in GIS software and all the routes within the area of interest were redrawn. This network was then validated and modified by the location of ancient ports, the AWMC map and geometries of islands. Afterwards, we were able to use network analysis to calculate centrality values in order to approximate attractiveness of harbours.

As strategic positioning is not the only determinant of trade intensity, we also decided to estimate the ratio of agricultural potential and the number of inhabitants on each island. This can then be used as an approximation of the urgency of goods import. Food production was measured based on the geographical environmental model we created. Various approaches to population estimates were put forward and discussed. We decided to use historical censuses from the 19th century, as they were taken before the demographic transition in the area and the population values should more or less correlate.

In order to validate the hypothesis emphasizing the Ptolemaic political actions as a key factor of the spread of the cults, we also needed to define the political factors. The Ptolemaic garrisons dispersed in the Aegean Sea are of great importance in our research because many of them were maintained for a long period of time and could thus increase the chance of spreading the cults via Ptolemaic soldiers. The second potential factor influencing the spread of the Egyptian cults is the existence of political leagues. In the context of the Aegean Sea, Ptolemies led the Island league from ca. 287 to 250s BCE. Again, this political factor is relevant because it is geographically delimited and lasted for a longer period of time.

The final and perhaps the most important dataset includes the locations of the Egyptian temples and artefacts from the above specified spatio-temporal frame (see Figure 1), collected from the RICIS catalogue (Bricault, 2005).

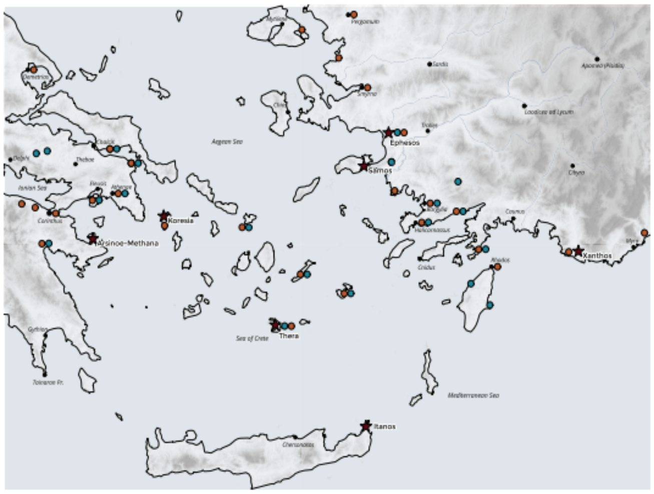


Figure 1: Map of the Aegean Sea region and locations of relevant temples (blue dots) and artefacts (red dots) of Egyptian cults, as well as Ptolemaic military garrison placements (stars).

Geographical analysis was used to calculate distances from each island to the closest army base and temple in order to evaluate spatial dispositions.

The final part of the research is the statistical analysis of data obtained in previous steps. At that point we had a table containing the list of islands with their attributes based on the data (e.g. if an island occupied a strategic position on the network, or if it had a Ptolemaic garrison) and the goal was to find a solid and interpretable mathematical model that would be able to find and explain relations between these values, mainly their dependency on the distance to the closest temple on the network.

Results and Conclusions

The mathematical model we selected using the statistical analysis of covariance uncovered a few patterns within our datasets. The results suggest that there is a strong correlation between the placement of the Ptolemaic garrisons and the dissemination of Egyptian temples and artefacts in the Aegean sea. The interpretation would, in this case, be that the Egyptian military forces potentially played a significant role in the spread of the cults. However, the analysis also showed that in areas far from these military bases the islands with higher trade attractiveness intensified the presence of the Egyptian cults.

The main task of this paper is to show how we can "regrow" the Egyptian cults in space and time using innovative computational methods from multiple disciplines. This allows us to validate the selected hypotheses from the academic discussion constructed by using traditional historiographical methods. We see great potential for using these approaches in the other case studies of GEHIR, as well as in historiography in general.

Bibliography

Arnaud, P. (2005). *Les routes de la navigation antique: itinéraires en Méditerranée*. Editions Errance.

Bricault, L. (2005). *Recueil des inscriptions concernant les cultes isiaques: RICIS*. Paris: Académie des inscriptions et belles-lettres.

Cumont, F. (1911). *The Oriental religions in Roman paganism*. Chicago: University of Chicago Press.

Fraser, P. M. (1960). *Two Studies on the Cult of Serapis in the Hellenistic World*. CWK Gleerup.

Rivers, R., Knappett, C., and Evans, T. (2013). "What makes a site important? Centrality, gateways and gravity." In Knappett, C. (ed), *Network Analysis in Archaeology: New Approaches to Regional Interaction, Oxford University Press, Oxford*, pp.125-150.

Turchin, P., Currie T. E., Turner, E. A. L. and Gavrilets, S. (2013). War, space, and the evolution of Old World complex societies. *Proceedings of the National Academy of Sciences*, *110*(41), pp.16384-16389.

Variation in academics' conceptions of e-assessment

Mike Mimirinis
mike.mimirinis@anglia.ac.uk
Anglia Ruskin University (Cambridge), United Kingdom

Earlier research in academic development proposed that enhancing teaching often derives from changes in how teachers think about their own teaching (Dall'Alba, 1991). Accordingly, the current study postulates that academic teachers conceptualise e-assessment in a number of qualitatively different ways; some of them may replicate traditional pedagogical approaches and merely transfer them to blended or online assessment settings while others may endorse transformational notions of assessment, underpinning knowledge construction and student agency. The study aims to synthesize and extend the findings from two existing clusters of studies. The first cluster derives from qualitative, non-phenomenographic research exploring conceptions held by academics about assessment without taking into account the layer of complexity brought about by technological tools. Influential studies in teachers' conceptions of teaching and learning identified a continuum of conceptions ranging from teacher-focused and content-oriented conceptions on the one end to student-focused and learning-oriented on the other end (for a review see Kember, 1997). Entwistle (2000) noted that *limited* evidence suggested that contrasting conceptions of teaching tend to hold corresponding views on assessment. Several studies also contended that teachers' conceptions of learning and teaching affect their approaches to teaching (e.g. Kember & Kwan, 2000). However, there have been no extensive, follow-up studies aimed at ascertaining whether contrasting conceptions of teaching are aligned to corresponding views on assessment or what the exact nature of such an alignment may be. Samuelowicz and Bain (2002) identified a continuum ranging from an emphasis on knowledge reproduction to an emphasis on knowledge construction and transformation and Postareff et al. (2012) described categories of conceptions, from reproductive conceptions to more transformational conceptions of assessment. The second cluster consists of phenomenographic studies in the area of conceptions of teaching and learning through technologies. These studies investigated university teachers' conceptions of, and approaches to e-learning (e.g. Ellis et al., 2009; González, 2010) and blended learning (Ellis et al., 2006). They did not, however, take into account assessment as a distinct element of the process of university teaching.

The adopted phenomenographic approach aimed to describe the phenomenon from the perspective of people involved with the phenomenon (Marton & Booth, 1997) i.e. teachers e-assessing within blended and online environments. Twenty participants have been invited to attend semi-structured interviews - preceded by two pilot interviews. The sample was drawn from teachers of a modern British university; a wide number of disciplinary backgrounds and academics with differing levels of experience was sought. The interviewers prompted academics to describe *possible* ways of utilizing e-assessment tools, their experiences of practicing e-assessment in general and details about a particular e-assessment practice they are engaged in. Participants were invited to explain what the purpose of the e-assessment was, how they understood their role in e-assessment and what they believed the role of the student was. Rounds of iterative analysis produced four qualitatively different, hierarchically-inclusive and logically-related categories of how academics conceive of e-assessment. E-assessment has been seen as a means: (i) managing and streamlining the assessment process (ii) communicating and engaging with students (iii) enhancing student learning and the quality of teaching (iv) community and (digital) identity building. The paper also reports on dimensions of variation, most importantly the role of the teacher/assessor, the role of the student, the level and quality of collaboration, the role of the technological medium and, finally, who benefits from the e-assessment processes. Conclusively, variation in the way academics conceive of e-assessment is discussed in light of previous studies in this area; implications for faculty development are approached under the lens of expanding awareness i.e. how can faculty development promote qualitatively more advanced conceptions of e-assessment and how can such interventions can improve student learning.

Bibliography

Dall' Alba, G. (1991). Foreshadowing conceptions of teaching. *Research and Development in Higher Education*, *13*, 293–297.

Ellis, R. A., Hughes, J., Weyers, M., & Riding, P. (2009). University teacher approaches to design and teaching and concepts of learning technologies. *Teaching and Teacher Education*, *25*(1), 109–117.

Ellis, R., Steed, A., & Applebee, A. (2006). Teacher conceptions of blended learning, blended teaching and associations with approaches to design. *Australasian Journal of Educational Technology*, *22*(3), 312–335.

Entwistle, N. (2000). Promoting deep learning through teaching and assessment: conceptual frameworks and educational contexts. Presented at the TLRP Conference, Leicester. Retrieved from http://www.etl.tla.ed.ac.uk//docs/entwistle2000.pdf

González, C. (2010). What do university teachers think eLearning is good for in their teaching? *Studies in Higher Education*, *35*(1), 61–78.

Kember, D. (1997). A reconceptualisation of the research into university academics' conceptions of teaching. *Learning and Instruction*, *7*(3), 255–275.

Kember, D., & Kwan, K.-P. (2000). Lecturers' approaches to teaching and their relationship to conceptions of good teaching. *Instructional Science*, *28*(5), 469–490.

Marton, F., & Booth, S. A. (1997). *Learning and Awareness*. Mahwah, NJ: L. Erlbaum Associates.

Postareff, L., Virtanen, V., Katajavuori, N., & Lindblom-Ylänne, S. (2012). Academics' conceptions of assessment and their assessment practices. *Studies in Educational Evaluation*, *38*(3–4), 84–92.

Samuelowicz, K., & Bain, J. D. (2002). Identifying academics' orientations to assessment practice. *Higher Education*, *43*(2), 173–201.

eLexicon. Dictionary of Polish Medieval Latin: from TEI encoding to an eXist-db application

Krzysztof Nowak
krzysztof.nowak@ijp-pan.krakow.pl
Polish Academy of Sciences Krakow, Poland

Introduction

From the *Lexicon* to the *eLexicon*

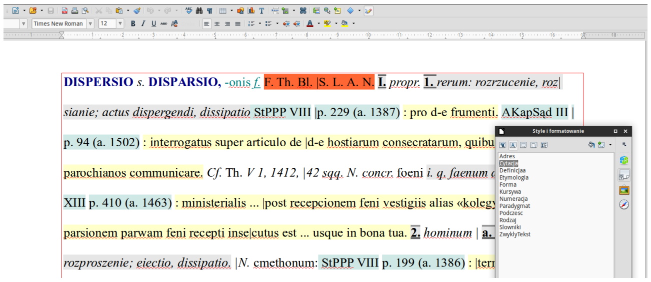
The first fascicle of the *Lexicon mediae et infimae Latinitatis Polonorum* (Dictionary of Polish Medieval Latin, henceforth LMILP) was published in 1953. The project aims at providing an exhaustive account of the Latin vocabulary used on Polish territory during the Middle Ages. Addressed to a scholarly public, the dictionary does not make many concessions to a less advanced user. Information is often conveyed only indirectly, by means of typographic devices or is left to be inferred by the reader. The project of retro-digitization of the LMILP started in the mid-2011 and was completed by the mid-2014. The web application, although completed, is still subject to modifications and refinements.

Dictionary Annotation

The XML encoding of the dictionary was by no means an ultimate goal of the project. Instead, the idea was to make the rich content of the LMILP fully searchable through a user-friendly interface. This objective, however, has deeply influenced the XML schema design. The TEI (TEI Consortium 2016) was chosen as an annotation standard because at the time the work started it had been already employed in major electronic lexicography projects (Lewis-Short by the Perseus Project; DuCange by the ENC). The popularity that the standard had gained among scholars contributed to emergence of lively community which produced documentation and use cases which supplemented the "Dictionaries" chapter of the TEI Guidelines. Also, the very fact that the TEI Guidelines offered a set of ready-to-use tags for the description of lexicographic content was not without significance. Finally, the TEI XML was supported by major software providers, an important factor for the project in which adaptation of existing rather than writing new software was planned.

Workflow

The paper dictionary was scanned and the output of the OCR program (Abbyy FineReader 11) was exported to ODT files; from each a *content.xml* file was extracted and then applied a series of XSL transformations. The main goal was to simplify styles that were automatically generated by the OCR software. Resulting XML files underwent second phase of XSL processing in which constitutive parts of the dictionary, such as entry, headword, sense definition *etc.*, were encoded. The output XML files were again re-translated into ODT format: entries were encoded as paragraph styles, other tags were represented as character styles. In the next phase of the project the lexicographers started to proofread OCR text and correct errors of automatic annotation. This task was performed with the help of LibreOffice Writer exclusively without annotators being actually conscious of the underlying XML structure. From the practical point of view, annotation consisted in verifying whether automatic XSL processing produced correct styles; if this was not the case correct style had to be applied, as in standard text processing task.



DISPERSIO *s.* **DISPARSIO,** -onis *f.* F. Th. Bl. |S. L. A. N. **I.** *propr.* **1.** *rerum: rozrzucenie, roz|*
sianie; actus dispergendi, dissipatio StPPP VIII |p. 229 (a. 1387) : pro d-e frumenti. AKapSąd III |
p. 94 (a. 1502) : interrogatus super articulo de |d-e hostiarum consecratarum, quibu
parochianos communicare. *Cf.* Th. *V 1, 1412, |42 sqq. N. concr.* foeni *i. q. faenum*
XIII p. 410 (a. 1463) : ministerialis ... |post recepcionem feni vestigiis alias «kolegy
parsionem parwam feni recepti inse|cutus est ... usque in bona tua. **2.** *hominum* | **a.**
rozproszenie; eiectio, dissipatio. |*N.* cmethonum: StPPP VIII p. 199 (a. 1386) : |terr

Fig. 1: "XML-unaware" annotation in the LibreOffice window

This approach allowed for reducing the learning curve to a minimum so that the team members could focus on the lexicographic content. However, it also has a serious drawback: annotation in the text editor cannot produce more complex hierarchies, since paragraph and character styles allow for representing at best two levels deep nesting.

TEI for the *eLexicon*

A guiding principle of the subsequent TEI annotation was to combine *editorial* and *lexical view* of the dictionary content by (1) preserving its original text and (2) storing normalized data in attributes and empty XML elements. Typographic properties of the text, on the other hand, were not generally encoded, they are easily reconstructible though.

Automatic and manual annotation consisted in three major procedures:

a. translation: custom ODT styles (corresponding to elements of dictionary structure) were "translated" into respective TEI elements or attributes;
b. grouping: deeply nested XML structure was produced from flat annotation;
c. enrichment: implicit information was made explicit.

Translation

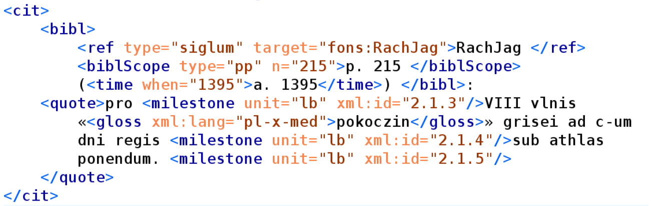
The paper justifies some of the annotation choices. Special attention is given to the peculiarities of encoding a scholarly lexicographic work.

1. **<entryFree>** element was chosen as a container for dictionary entries.
2. Essential features of the dictionary macro- and microstructure are encoded as: **<form>, <orth>; <gramGrp>, <gen>, <iType>, <pos>; <etym>, <lang>, <mentioned>; <cit>, <bibl>, <biblScope>, <date>, <quote>; <sense>, <usg>, <def>, <gloss>; <xr>, <ref>; <lbl>; <re>, <certainty>, <oVar>, <note>.**
3. Content and form peculiarities of the LMILP are reflected in respective attributes. So, for example, functional variation of the entries is represented in the **@type** attribute of the **<entryFree>** and can take one of the following values: **main, xref, hom.**
4. The TEI schema was only lightly customized: unused elements were deleted; a few content restrictions were overcome.

Grouping: adding depth

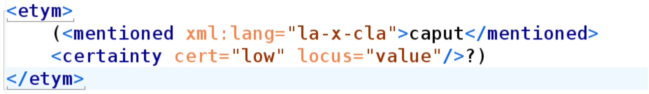
The flat entry structure had to undergo heavy XSL processing, so deep nesting typical of scholarly dictionaries could eventually emerge. Relative ease of the XML-unaware manual annotation resulted in time-consuming post-processing. The **xsl:for-each-group** XSL function was employed in order to structure:

1. citation groups:



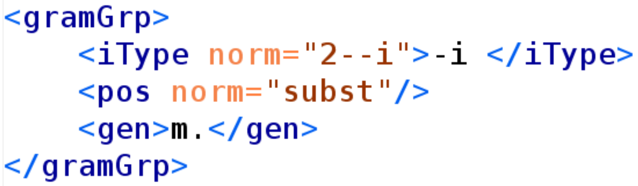
```
<cit>
    <bibl>
        <ref type="siglum" target="fons:RachJag">RachJag </ref>
        <biblScope type="pp" n="215">p. 215 </biblScope>
        (<time when="1395">a. 1395</time>) </bibl>:
    <quote>pro <milestone unit="lb" xml:id="2.1.3"/>VIII vlnis
        «<gloss xml:lang="pl-x-med">pokoczin</gloss>» grisei ad c-um
        dni regis <milestone unit="lb" xml:id="2.1.4"/>sub athlas
        ponendum. <milestone unit="lb" xml:id="2.1.5"/>
    </quote>
</cit>
```

2. etymological groups:



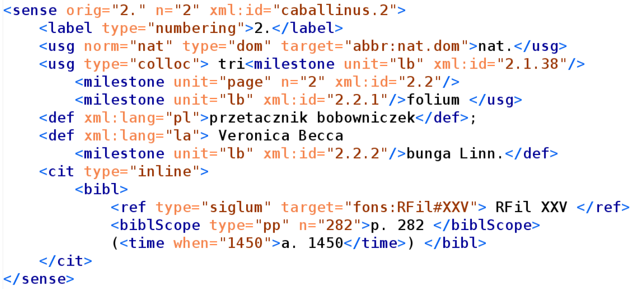
```
<etym>
    (<mentioned xml:lang="la-x-cla">caput</mentioned>
    <certainty cert="low" locus="value"/>?)
</etym>
```

3. PoS and grammar groups:



```
<gramGrp>
    <iType norm="2--i">-i </iType>
    <pos norm="subst"/>
    <gen>m.</gen>
</gramGrp>
```

4. sense groups:



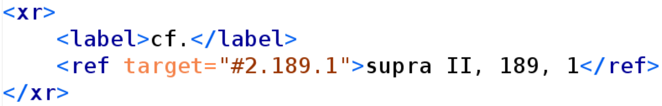
```
<sense orig="2." n="2" xml:id="caballinus.2">
    <label type="numbering">2.</label>
    <usg norm="nat" type="dom" target="abbr:nat.dom">nat.</usg>
    <usg type="colloc"> tri<milestone unit="lb" xml:id="2.1.38"/>
        <milestone unit="page" n="2" xml:id="2.2"/>
        <milestone unit="lb" xml:id="2.2.1"/>folium </usg>
    <def xml:lang="pl">przetacznik bobowniczek</def>;
    <def xml:lang="la"> Veronica Becca
        <milestone unit="lb" xml:id="2.2.2"/>bunga Linn.</def>
    <cit type="inline">
        <bibl>
            <ref type="siglum" target="fons:RFil#XXV"> RFil XXV </ref>
            <biblScope type="pp" n="282">p. 282 </biblScope>
            (<time when="1450">a. 1450</time>) </bibl>
    </cit>
</sense>
```

Enrichment: expanding the dictionary content

Considerable effort was put into enriching the original content of the dictionary, namely: (1) resolving references, (2) normalizing strings, (3) adding redundant and/or inferred information.

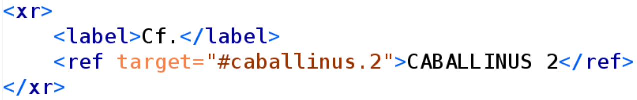
Resolving references

A typical reference to an exact location in the dictionary text was encoded as follows:



```
<xr>
    <label>cf.</label>
    <ref target="#2.189.1">supra II, 189, 1</ref>
</xr>
```

References to a specific entry or sense relied on the **@xml:id** attribute:



```
<xr>
    <label>Cf.</label>
    <ref target="#caballinus.2">CABALLINUS 2</ref>
</xr>
```

The encoding of most frequent type of references (pointing to a source of a language use example) is illustrated in the section II B 4 above.

String normalization

By string normalization, we mean a set of various procedures applied in order to generate a *lexical view* of the dictionary content. Standardized strings are usually stored in **@norm** attributes of such elements as language or usage labels, prepositional and inflectional patterns *etc.* Their primary goal is to enable unified search that would be agnostic of the exact formulation of the paper dictionary. For example, when looking up philosophy-related terms one should be able to retrieve them no matter whether they have been marked with a *phil.* label or with more verbose formula *in textibus philosophicis* "in philosophical texts", as both are annotated as **@norm="phil".** The second major goal of the normalization was to render chronological information consistent and machine-readable. Its proper annotation should reflect the fact that many medieval texts cannot be dated but only approximately. Apart from some obvious cases (`@when` attribute stores a year date, for example **<time when="1450">a. 1450</time>)** the LMILP employs:

1. century dates **(<time notBefore="1401" notAfter="1500">saec. XV</time>)**
2. imprecise dates in year **(<time notAfter="1120">ante 1120</time>**) or century format (**<time notBefore="1401" notAfter="1450">saec. XV in.</time>**).

Making information explicit

Finally, substantial effort has been devoted to making explicit what is not expressed directly in the paper dictionary, but left to be inferred by an expert user. In the LMILP, this is the case, for example, of a part of speech label which is provided for adverbs or conjunctions, but is normally omitted from verb or noun entries. Empty elements have therefore been created and their attributes filled with the inferred content. So, in a typical case, an element **<pos norm="subst"/>** would be appended to a **<gramGrp>** group whenever the paper dictionary informs about a word's part-of-speech only indirectly, by means of a gender label (*f.* for Lat. *femininum*) or inflectional ending typical of nouns (*-ae*):

The Dictionary Web Application

The last part of the paper briefly presents the overall architecture of the dictionary web application, its user interface having been already described elsewhere (Nowak 2014). Written entirely in XQuery, the application is served directly from the eXist-db instance with HTML and JavaScript code being equally stored in a database or generated on the fly. The presentation focuses on those features available in the eXist-db which are of critical importance for dictionary application design:

1. Various types of indexes available in the eXist-db allow for efficient retrieval of content from deeply nested dictionary files and dispersed textual data.
2. A templating system allows for fine-grained web presentation of the XML content.
3. A URL rewriting engine supports a logical system of dictionary content access.
4. An out-of-the-box RESTful API exposes lexicographic content to external applications.

In the conclusion, I will also point to some difficulties that I have encountered and which have mainly to do with handling application's state, a crucial feature for multi-language tools which require storing user search results.

Bibliography

Nowak, K. (2014). ‘The eLexicon Mediae et Infimae Latinitatis Polonorum. The Electronic Dictionary of Polish Medieval Latin’. In *The User in Focus. Proceedings of the XVI EURALEX International Congress: 15 - 19 July 2014, Bolzano/Bozen*, edited by Andrea Abel, Chiara Vettori, and Natascia Ralli, 793–806. Bolzano: EURAC Research.

TEI Consortium. (2016). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version: 3.0.0. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Uncovering ‘hidden’ contributions to the history of Digital Humanities: the *Index Thomisticus*’ female keypunch operators

Julianne Nyhan
julianne.nyhan@gmail.com
University College London, United Kingdom

Melissa Terras
m.terras@ucl.ac.uk
University College London, United Kingdom

Introduction

Who undertook the foundational work of the discipline now known as Digital Humanities (DH)? Whose work merits inclusion in the history of the genesis of DH - the leaders of scholarly projects? Their research assistants? Their administrators? Their funders? Have important contributions to early DH projects gone unacknowledged or been silenced by the field’s dominant ‘founding father’ narratives? How can a better understanding of previously undocumented contributions to the founding of DH allow us to evaluate the centrality of processes like collaboration and interdisciplinarity to the development and establishment of DH?

This paper describes our research on the ‘hidden contributions’ to the *Index Thomisticus* project of Fr Roberto Busa S.J. (1913-2011). Busa is often said to be the founding father of DH: “Most fields cannot point to a single progenitor, much less a divine one, but humanities computing has Father Busa, who began working (with IBM) in the late 1940s on a concordance [the *Index Thomisticus*] of the complete works of Thomas Aquinas” (Unsworth 2006; see also Hockey 2004 and the more nuanced analysis Busa’s role in Jones (2016)). Our research has uncovered the details and nature of the contributions made by the punched card operators who transcribed the (pre-edited) texts of Thomas Aquinas and related authors into machine-actionable data using punched card technology, thus completing the essential preliminary work on the *Index Thomisticus*. The operators were the mostly female trainees of the keypunch school that Busa had set up in Milan in 1956 (and that ran until c.1967) as well as the female keypunch operators who worked with him in his Literary Data Processing Centre (CAAL). In addition to recovering the specifics of their work we have also sought to better understand their personal experiences of working on the project and whether the skills they learned were of subsequent benefit to them. Despite the formidable amount of work that they undertook, and the crucial nature of their task to the project, the identities of these women and the nature of their contributions were largely unknown and unacknowledged until this research was undertaken.

Methodology

Previous research on the history of DH has shown that when used with care oral history can contribute to a grounded history that exposes overarching processes while acknowledging through personal narratives the agency and creativity of a plurality of individuals, and not just the great men and women of scientific advancement (Nyhan, Flinn, Welsh 2015). An oral history approach was again adopted for this project; ten of the female punched card operators who had worked with Busa for various durations between 1954 and 1970 were interviewed.

The interviews were carried out from April 1st to 3rd 2014 in the Alosianum College of Philosophical Studies, Gallarate, Italy. Nyhan was present throughout though the interviews were carried out in Italian by Marco Passarotti (a former student of Busa and Principal Investigator of CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan) as Nyhan does not speak Italian. Passarotti was given a set of core questions in advance that drew heavily on Nyhan's wider research for the Hidden Histories project (see, for example, Nyhan and Flinn 2016). The questions were prepared with two aims in mind: that of uncovering the Womens' memories of working on the *Index Thomisticus* project as well as capturing their memories of Busa himself. The core questions were:

- How did you first hear about punched card technology?
- How did you find out about the punched card school?
- How did you secure a place there?
- What training did you receive?
- What kinds of work did you do?
- Was the training you received useful to you in later life?
- Please share a memory of Fr Busa?

It was also agreed that Passarotti should ask other questions as he saw fit, for example, in response to an interesting point that was raised by an interviewee. All interviewees signed a waiver form in advance of the interview that gives permission for their recollections to be published. A grant from the European Association for Digital Humanities (EADH, see the call for funding proposals) was secured so that the recorded interviews could subsequently be transcribed and translated by a Research Associate (Ana Vela), who is fluent in both Italian and English. Nyhan then worked through the translations in order to edit them further for clarity and check them, as far as was possible, for factual accuracy. She subsequently carried out a close reading and qualitative analysis of the interviews in order to identify common themes and telling divergences. This was followed by a historical-interpretative analysis that compared and contrasted the issues identified in the oral history sources with relevant primary and secondary literature. Finally, we wrote the results up as narrative history.

Findings

What emerges from the interviews is an insight into the social, cultural and organisational conditions that the female punched card operators worked under and how they were treated in what was, structurally at least, a male-dominated environment. The interviews contain a wealth of recollections about the following issues in particular:

- The womens' discovery of the training school that Busa set up
- Their entrance test to the school
- Their training and tasks as keypunch operators
- The organisational hierarchy of the *Index Thomisticus* workforce
- Their awareness of the aims of the project and of Humanities Computing and Computational Linguistics more generally
- Their knowledge of Latin
- Usefulness of the training to them in later life
- Their memories of Busa.

For example, regarding the usefulness of their training, it opened opportunities that would otherwise have been blocked to them. A number of them went on to work as keypunch operators on an early machine translation project in the EURATOM Center at Ispra, Milan (On Busa's connection to Ispra, see Busa, 1980, p.86). Nevertheless, the interviews collectively give the sense that the women were seen as a source of low-cost and low-skilled labour. They did not have opportunities to progress from the position of keypunch operator and their training seems to have been the minimum necessary to carry out their roles. Most were not even made aware of the wider significance or aims of the *Index Thomisticus* project. Despite the existence of other research projects like

EURATOM, mentioned above, and that an 'employment path' in the context of research computing was beginning to open up, their potential longer-term contributions to such work were not considered or fostered.

Conclusion

It is almost a cliché to say that DH's collaborative nature makes it distinct and differentiates it from traditional Humanities. However, our research on the *Index Thomisticus* project has prompted us to ask whether claims about the centrality of collaboration to DH are more problematic than they first appear. As we will show, collaboration was the basis on which Busa's *Index Thomisticus* was realised. However, in the 'incunabular phase' (see Rockwell et al. 2011) of DH some forms of collaboration were considered more worthy than others and the contributions of the many female (and occasionally male) punched card operators who did the work of the project were not acknowledged. Until our research, their identities and the nature of their contributions had essentially disappeared from both the historical record and the collective memory of the DH community. This gives rise to a number of interrelated questions that have not yet been adequately addressed by scholarship on the history of DH: when and how did collaboration take on its significance for the field? What has influenced decisions about what kinds of DH collaborations have and have not tended to be acknowledged and how has this changed over time? What is the significance of the alleged cleaving of DH from the practices of the mainstream Humanities in regard to collaboration? Accordingly, our paper will also aim to open a wider discussion about the history of collaboration and the role it played in the formation and establishment of DH.

Bibliography

Busa, R. (1980). "The Annals of Humanities Computing: The *Index Thomisticus*." Computers and the Humanities 14 (2): 83–90

Jones, S. E. (2016). *Roberto Busa, S. J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. Routledge.

Hockey, S. M. (2004). "The History of Humanities Computing." In *Companion to Digital Humanities (Blackwell Companions to Literature and Culture)*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Blackwell.

Nyhan, J., Flinn, A. & Welsh, A., (2015). "Oral History and the Hidden Histories Project: Towards Histories of Computing in the Humanities." *Digital Scholarship in the Humanities* 30(1): 71–85. Available at http://dsh.oxfordjournals.org/content/30/1/71/. First published online in *Literary and Linguistic Computing*, July 30, 2013.

Nyhan, J and Flinn, A. (2016). *Computation and the Humanities: towards and Oral History*. Springer.

Rockwell, G., Smith, V., Hoosein, S., Gouglas, S., and Quamen, H. (2011). "Computing in Canada: A History of the Incunabular Years." In Digital Humanities 2011: conference abstracts. Stanford University Library. Pp. 207–10.

Unsworth, J. (2006). "Digital Humanities Beyond Representation." University of Central Florida, Orlando, FL, November 13, 2006. http://people.brandeis.edu/~unsworth/UCF/.

Trauma Representations in Videogames

Jennifer Olive
jolive1@gsu.edu
Georgia State University, United States of America

Over the last few years, the videogame community has seen an increase in the visibility of indie games, particularly ones that focus on serious subjects. Given the name "empathy games" by Minority Media's Vander Caballero in his 2014 GDC talk, this sub-genre is known for its intense and meaningful narratives that often bring to light traumatic experiences. While the classification of these games under empathy can be problematic as highlighted by Anna Anthropy (2015) and the responses to her game Dys4ia, the phenomenological experience of playing these videogames offers a new perspective through which one can engage understandings of new media aesthetics and narrative design. Moreover, Caballero's (2014) identification of the "empathy game" category as strategic for project support highlights the development of the sub-genre at a point which methodologies of analyzing these aesthetics and design choices as they relate to the representation of trauma within videogames are needed most. Building on the relevance of the issue as well as the unexplored intersections that such mediation brings to prior understandings of critical theory, this paper will propose a methodology for analyzing representations of trauma in videogames that builds on current work in the field. Moreover, it will explore the theoretical implications of such a method to the model of player experience through remediation in order to offer a broader understanding of the meaningful intersection of trauma narratives and new media.

Beginning with its most well-known applications in the 1990s with the work of Cathy Caruth (1995, 1996) in literary studies and in the 2000s with the work of Dominick LaCapra (2014) in history, the study of trauma has spread from its Freudian psychoanalytic roots to become a field of study in both the sciences and the humanities. Both of the models developed by and around these two theorists are extremely helpful for literary representations given their foundations in disciplines that largely rely on print media and are able to be applied to different genres given their multidisciplinary foci; however, the study of trauma representations has not yet spread widely to new media applications aside from its application in testimonial media. This exclusion suggests a gap in scholarship for trauma studies as it does not account for how these models of trauma may be useful in or influenced by other media.

With the increasing cultural and economic impact of videogames, it is essential that the digital humanities develop a critical framework for understanding and analyzing the intersections of trauma studies and videogames. Tobi Smethurst's work (2014, 2015a, 2015b) attempts to bridge this gap by identifying how both game studies and trauma studies could benefit from an interdisciplinary application. Smethurst's doctoral thesis (2015a), explores this intersection to offer a tripartite methodology for analyzing trauma in videogames: interreactivity, empathy, and complicity. This methodology is particularly useful because it examines the rhetorical feedback loop between player and game, which, she argues, creates a very specific experience for representing trauma by implicating the player as a perpetrator, and it works very well for games that are designed to elicit such feedback from players. This methodology, however, does not fully account for videogames being able to represent trauma in ways that do not end in making the player complicit in the trauma.

Expanding on this foundation, my paper will discuss the use of videogames as an interactive digital media platform for representing traumatic narratives, identify a methodology for analyzing such representations, and argue for a model that understands play in videogames containing representations of trauma as a form of witnessing. This work will begin by grounding its understanding of trauma as it is defined in literary studies through the works of Caruth (1995, 1996) and LaCapra (2014) respectively and expanding its theoretical framework to the use of digital media through digital narrative and media studies by exploring the affordances of videogames highlighted by Janet Murray (1997) in Hamlet on the Holodeck. Next, I will discuss a methodology for analyzing representations of trauma in videogames that utilizes those media aesthetics and the object's use of procedural rhetoric as defined by Ian Bogost (2007) in Persuasive Games to create a remediated (Bolter and Grusin, 2000) experience of witnessing as described by Dori Laub (1995) in "Truth as Testimony: The Process and the Struggle." In doing so, I will discuss how this methodology provides a much needed perspective concerning the intersection of videogames and trauma studies. Additionally, I will argue for a model of representing trauma in videogames that seeks to understand feedback loops with the audience in broader terms, which I will ground in the understanding of videogame ethics as discussed in The Ethics of Computer Games (2009) by Miguel Sicart. Such a model will focus on the mediation of the trauma through the videogame experience and account for the overlapping of ethical systems that exist for the player outside of the gameworld, which, I will argue, helps the player navigate the different levels of witnessing. Together, this combination of theoretical foundations will expand on the current scholarship to offer a more holistic understanding of the potential held by videogames as a medium to represent trauma.

Bibliography

Anthropy, A. (2015). "Empathy Game." Web. http://auntiepixelante.com/empathygame/

Bolter, J., and Grusin, R. (2000). Remediation: Understanding New Media. Cambridge, MA: The MIT Press

Caballero, V. (2014). "Empathy Games." GDC 2014. Web. http://www.gdcvault.com/play/1020598/Empathetic-Games-Are-Here-to

Bogost, I. (2007). Persuasive Games: The Expressive Power of Videogames. Cambridge, MA: MIT Press.

Caruth, C. (Ed.). (1995). Trauma: Explorations in Memory (pp.61-75). Baltimore, MD: Johns Hopkins University Press.

Caruth, C. (1996). Unclaimed experience: Trauma, Narrative and History. Baltimore, MD: Johns Hopkins University Press.

LaCapra, D. (2014). Writing History, Writing Trauma. Baltimore, MD: Johns Hopkins University Press.

Laub, D. (1995). "Truth as Testimony: The Process and the Struggle." In C. Caruth (Ed.), Trauma: Explorations in Memory (pp.61-75). Baltimore, MD: Johns Hopkins University Press.

Murray, J. H. (1997). Hamlet on the Holodeck: The Future of Narrative in Cyberspace. New York, NY: The Free Press.

Sicart, M. (2009). The Ethics of Computer Games. Cambridge, MA: The MIT Press.

Smethurst, T. and Craps, S. (2014). "Playing with Trauma Interreactivity, Empathy, and Complicity in The Walking Dead Video Game." Games and Culture, 10, 3, 269-290.

Smethurst, T. (2015a). Playing with trauma in video games: interreactivity, empathy, perpetration (Doctoral dissertation). Ghent University, Belgium.

Smethurst, T. (2015b). "Playing Dead in Videogames: Trauma in Limbo." The Journal of Popular Culture, 48,5, 817-835.

Many Tongues, Multiple Networks: A Corpus-Based Pragmatic Study of Multilingualism on Nigerian Virtual Political Space

Tunde Olusola Opeibi
bopeibi@unilag.edu.ng
University of Lagos, Nigeria

Olusola Abiodun Aina
solar_ai2003@yahoo.co.uk
Crawford University, Nigeria

This study examines the increasing utilisation of the Internet and related digital media technologies such as Facebook, Twitter and Youtube for political activities and civic engagement in Nigeria. It however focuses specifically on the ways in which political actors and citizens in Nigeria draw on the existing multilingual resources within the Nigerian sociolinguistic ecosystem in their web-based political interactions.

The study also discusses how these stakeholders use a combination of mixed codes in online interactions with political issues and political stakeholders to reshape the political space and promote participatory democracy.

Theoretical insights that underpin the study adopt approache from **computer-mediated discourse analysis** and **pragmatics.** The data set consists of a range of media multilingual-based online political posts between 2011 and 2015 elicited from our corpus construction project, CONNMDE (Corpus of Nigeria New Media Discourse in English). The methodological procedures also utilise some digital humanities software packages and concordancers such as Sketch Engine and AntConc3.4 for data harvesting and to complement the qualitative and quantitative analyses (Baker, et al, 2008:275).

Generally, the study reports some findings on the ongoing digital humanities project that focuses on new media corpus construction. It then isolates and identifies clear instances of deliberate deployment of multilingual resources influenced by and rooted in ethnolinguistic identities as communication inclusion strategy. It equally confirms that Nigerian politicians use political websites with multilingual texts to encode acts of political persuasion and convey pragmatic meanings.

This presentation concludes by suggesting that online multilingual discursive practices may be indicative of or closely linked to people's purpose-driven offline sociolinguistic practices and identities deployed for socio-pragmatic goals.

Modeling Creativity: Tracking Long-term Lexical Change

Peter Organisciak
organis2@illinois.edu
University of Illinois, United States of America

Samuel Franklin
samuel_franklin@brown.edu
Brown University, United States of America

The concept of creativity underwent a period of shifting meaning and rapid adoption in the twentieth century. Following from a narrow early scope of usage, in which it carried largely religious connotations, the word 'creative' grew broader and adopted the more subjective meanings we are familiar with today. Though many contemporary observers point out the vagueness of the term, creativity's power comes from a particular mix of meanings and connotations accrued over time. Still, there is no clear inventory of the higher-level concepts around discussion of creativity and how they evolved. Additionally, because of the rapid increase in usage, early uses of 'creativity' may be overlooked as they are overshadowed by much more common later uses.

In this paper, we present a method for tracking the different styles of discourse around a concept over time, developed for following the evolution of 'creativity' but applicable to other domains. Our approach is an application of Latent Dirichlet Allocation (LDA) -trained topic models, with three novel steps in their preparation:

- a highly-selective keyword sampling of pages from a large text corpus,
- temporally weighted training sample ordering, and
- purposively-assigned asymmetric document-topic priors.

Motivation

This research supports a larger project on the discourse of 'creativity' in post-WWII America. The anecdotal observation that creativity has become a buzzword in recent years is supported by graphs of word frequency available through platforms such as the Google Ngram viewer and JSTOR Data for Research, which show creativity only entered the American lexicon in the twentieth century, diffusing rapidly after about 1950. 'Creative' appears to have enjoyed a similar growth spurt over the same period, but it preceded creativity by about three hundred years.

Unfortunately, these graphs do not reveal the long-term changes in meaning nor the distinct contexts in which the language of creativity accrued its contemporary salience. It is obvious from contemporary usage that the word 'creative' has a tangle of interrelated but distinct meanings, ranging from *generative* or *constructive* to *artistic* to *nonconformist*. These meanings are distributed unevenly over time and across communities of discourse. To understand why and through what routes creativity arose when it did, it will be essential to tease apart these various meanings of creative, and the contexts in which they have been strongest over the long term.

We believe topic modeling can help. First, it can help us identify and distinguish between the several discourses in which creative has been a keyword—for example in theology versus education versus psychology—whilst still reflecting the historically shifting connections and overlaps between those. Second, we can then apply those topics to only those texts containing the token 'creativity,' to reveal which of the discourses and meanings of 'creative' seem to be at work. By this process we can achieve a more granular picture of the creativity boom, helping us answer the basic question 'what do we talk about when we talk about creativity?'

Approach

Topic modeling enables us to observe more higher-level concepts than keyword searching and collocations would allow. Topic modeling depends on a certain class of mixed model clustering, but we believe that the two should not be conflated. The connotation of 'topic modeling' implies a qualitative interpretability. Surfacing what would be recognized as concepts is not solely a case of running a modeling algorithm on words from a text. Instead, it needs to be paired with a series of preparatory and parameterization steps tailored to the particular problem.

We developed a workflow for training better topic models to track a specific concept in a temporally-biased corpus. This involves standard pre-processing such as stoplisting words, but also contributes three novel steps: selective page-level sampling, weighted training, and explicitly imbalanced prior assumptions on how likely a document is to be reflected by each topic. The sampling helps focus the models on creativity, the weighted training counteracts temporal biases to retain older topics to surface, and the asymmetric priors help find more granular topics.

For a dataset cross-cutting published work broadly, we used a recent release of the HTRC Extracted Features Dataset (Capitanu 2016). The Extracted Features Dataset includes term counts for every page of 13.7m volumes in the HathiTrust Digital Library and benefits from a mostly indiscriminate digitization policy, allowing us to observe a term's usage in a wide spectrum of texts.

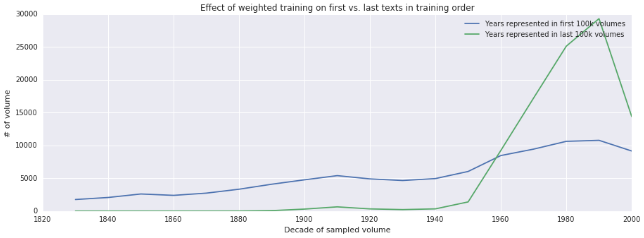
Topic Modeling Preparation

In topic modeling, the goal is surfacing patterns that represent qualitatively intuitive concepts. However, to the *methods* used for topic modeling, the mark of success is being able to represent documents in the desired number of topics with as little error as possible. This divergence between our needs and the machine's makes the text preparation important. One such preparation is to remove words that are not interesting to a human reader. An algorithm may find a meaning in a word like 'however' or 'whereas', but as a proxy for topicality, such words are usually not desired.

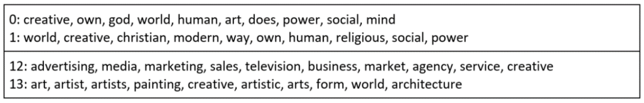
For tracking trends in creativity discourse, we used Latent Dirichlet Allocation (LDA) combined with standard preprocessing: removing the most common words in the English language, less interesting parts-of-speech (e.g. adverbs, determiners, numbers), and cutting off the sparser end of the vocabulary. In addition, we developed three less common preparations in the service of issues arising from tracking concept diffusion.

Sampling. One possible approach to finding the most common topics for a keyword is to look at the underlying term-topic probabilities for the keyword, post-training, and identifying the topics where the word is most common. This approach scales well to multiple keywords but provides low specificity for tracking them. Instead, we sampled only pages that use the word 'creativity' or variants of 'creative'. The size of the HTRC EF Dataset affords the small contextual window and selective sampling, as there were slightly more than 2 million volumes found that have at least a single mention of the keyword list.

Weighted training. When training topic models, earlier texts have an outsize influence on the topics that emerge. This is a problem for our use case, where we expected a topical shift alongside a steep increase in usage. A randomized training order would reflex later texts very strongly, at the risk of missing topics which are prominent in older texts. To counteract this, we applied weighting to the randomized training order, to soften the temporal bias without entirely removing is. When deciding on the next text to send to the training algorithm, texts are weighted for sampling with weight(decade) = 1/ n(decade). The following figure shows this weighting in action: at the important start of training, newer texts are only slightly more common. Since a disproportionate number of older texts are used early on, there are few left by the end of training.



Honeypot topics. As part of the estimation process for LDA topics, we have to formalize our best guess for how likely any given topic is to be assigned to a document. Past work has found value in allowing for these prior assumptions to be uneven - e.g. one topic can be considered more likely than another (Wallach, Mimno, and McCallum 2009). We found initial success with a heuristic intended to find many smaller trends in the collection by provided the first three topics the majority of the probability mass and dividing the remainder between the remaining topics. In qualitative comparisons with evenly distributed probabilities, we found that setting asymmetric priors in this way set traps to catch broadly common documents in predictable topics, while allowing other topics to surface more highly-specific topical hotspots.


0: creative, own, god, world, human, art, does, power, social, mind
1: world, creative, christian, modern, way, own, human, religious, social, power

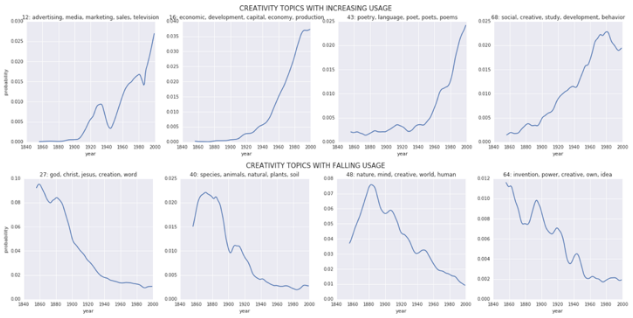
12: advertising, media, marketing, sales, television, business, market, agency, service, creative
13: art, artist, artists, painting, creative, artistic, arts, form, world, architecture

Two general topics and two niche topics

Results

The training yielded several topics which confirm where we would expect to find the language of creativity. Some of these reflect specialized uses, such as in advertising and evolutionary biology, while others reflect the broad humanistic discussions of the nature of thought, art, and religious creation. By graphing these topics over time we can see that our temporally weighted sampling appears to have been successful in revealing archaic topics that are nonetheless essential to understanding the connotative textures of the language of creativity in our own time.

The following figures show a small selection of topics where the usage has grown in the past 150 years, and topics where it has fallen. Generally, we see that the language of creativity has transitioned from religious and natural notions of creation toward the language of economic and human capital.



Future work

This work has a number of future directions. We have thus far focused on a number of words (creative, creativity, creativeness); moving forward, we intend to map how the verb and noun uses compare. Also, while much of the development has been qualitatively development against our particular problem, we hope to compare variants of our workflow in more contexts.

Conclusion

In the proposed paper, we will present our method for tracking longitudinal trends in a diffuse and shifting context. Motivated by work on the language of creativity and particularly the noun 'creativity', our contributions are in text processing and parameterization for topic modeling, allowing clear and specific concepts to be revealed.

Bibliography

Capitanu, B., Underwood, T., Organisciak, P., Cole, T. J., Sarol, M. J., Downie, J. S. (2016). *The HathiTrust Research Center Extracted Features Dataset.* 1.0. HathiTrust Research Center. Dataset. http://dx.doi.org/10.13012/J8X63JT3

de Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights.* Fordham University Press.

Wallach, H.M., Mimno, D.M., and McCallum, A. (2009). "Rethinking LDA: Why priors matter." *Advances in neural information processing systems.*

Williams, R. (1976). *Keywords: A Vocabulary of Culture and Society*. New York: Oxford University Press.

Building worksets for scholarship by linking complementary corpora

Kevin Page
kevin.page@oerc.ox.ac.uk
University of Oxford, United Kingdom

Terhi Nurmikko-Fuller
terhi.nurmikko-fuller@anu.edu.au
University of Oxford, United Kingdom

Timothy Cole
t-cole3@illinois.edu
University of Illinois, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois, United States of America

Background and General Motivation

The HathiTrust Digital Library

The HathiTrust Digital Library (HTDL) comprises digitized representations of 15.1 million volumes: approximately 7.47 million book titles, 418,216 serial titles, and 5.3 billion pages, across 460 languages. HTDL is best described as "a partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future".

The HathiTrust Research Center (HTRC) develops software models, tools, and infrastructure to help digital humanities (DH) scholars conduct new computational analyses of works in the HTDL. For many scholars the size of the HTDL corpus is both attractive and daunting: many existing DH tools are designed for smaller collections, and many research inquiries are facilitated by more focused, homogeneous collections of texts (Gibbs and Owens, 2012).

Worksets

In many, if not most, DH research endeavours, performing an analytical task across the whole HTDL is neither practical nor productive (Kambatla et al., 2014). For example, a tool trained to identify genre attributes of 18th century English language prose fiction may not be applicable to 20th century French poetry. The first step is to identify the subset -- of works, editions, volumes, chapters, pages -- to set an initial investigative scope and, indeed, subsequent iterative refinements of a subset as research proceeds. In a corpus as large and complex as the HTDL, finding materials and then defining the sought after subset can be extraordinarily difficult.

HTRC has come to call collections of digital items brought together by a scholar for her analyses a "workset", created to help the researcher build, manipulate, iteratively define and compare their collections. Reflecting upon input and advice from the DH community, Jett (2015) defines a workset as a machine-actionable research collection realised as:

1. An aggregation of members (volumes, pages, etc.);
2. Metadata intrinsic to the workset's essential nature (e.g., creator, selection criteria);

3. Metadata intrinsic to digital architectures (i.e. creation date & number of members);
4. Metadata supportive of human interactions (i.e. title & description);
5. Derivative metadata from workset members (e.g. format(s), language(s), etc.); and,
6. Metadata concerning workset provenance (e.g. derived from, used by, etc.).

Broadly, item 1 identifies the actual data used in an analysis; whereas the remaining metadata items describe the workset itself, aiding workset management throughout the research cycle.

Cross-corpus worksets

As alluded above, numerous criteria can be used to select the constituents of a workset; and several technological implementations could, in theory, realise worksets. In researching the design and realisation of worksets and associated tooling, we are also mindful to remain grounded in their practical application and the needs of scholarly users. We have therefore undertaken our work through discipline-based scenarios in which we can explore the strengths and weaknesses of the HTDL viewed through the prism of worksets.

We report one such exploration here, questioning *whether (relatively) small, well explored, and well understood corpora can be superimposed over the HTDL to aid navigation and investigation of the much larger and superficially understood HTDL collection?*

From a system perspective, a cross-corpus workset requires exposing *compatible* metadata (items 2-6 above) from multiple collections, first used to align common elements, and then to assemble worksets. We take a Linked Data approach and achieve compatibility through ontologies, which might initially be bibliographic (and derived from library records) but should be iteratively extensible into the domain of the subject of study.

Examples in early English print

Early English Books Online Text Creation Partnership (EEBO-TCP) is a partnership with ProQuest and over 150 libraries and universities, led by Michigan and Oxford, to generate highly accurate, fully-searchable texts tracing the history of English thought and learning from the first book printed in English in 1473 through to 1700. Between 2000-2009 EEBO-TCP Phase I converted 25,000 selected texts from the EEBO corpus into TEI-compliant, XML-encoded hand-transcribed texts, subsequently freely released in January 2015.

In the work reported here, we have conjoined EEBO-TCP with a HathiTrust subset consisting of all materials described in their metadata as being in English and published between 1470 and 1700.

To ensure a prototype which simultaneously explored the fit of scholars' needs to the technology and exercised the technical challenges outlined in the previous section, we undertook a 'complete circuit' through the datasets (Figure 1). We: (i) ran a consultative workshop to choose investigations which might form the basis of worksets; (ii) used these abstract worksets to identify concrete requirements for the conjoined metadata; (iii) generated metadata from both corpora according to these specifications; (iv) aligned elements from both datasets in an overlapping superset; (v) realised the worksets identified in (i) using this metadata.

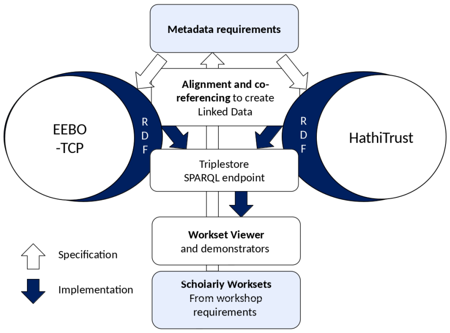

Metadata requirements
Alignment and co-referencing to create Linked Data
EEBO -TCP
R D F
R D F
HathiTrust
Triplestore SPARQL endpoint
Workset Viewer and demonstrators
Scholarly Worksets From workshop requirements
Specification
Implementation

Figure 1. Overview of the metadata circuit leading to our cross-corpora workset

Motivating worksets

Following the workshop we identified the following workset selections; we describe their implementation in subsequent sections:

- Find all the works, appearing in both datasets, written by Richard Baxter.
- Find works in both datasets published in Oxford.
- Find works published outside of London (where the bulk were published).
- Find works from both datasets published outside of London in the mid-to late 1600s.
- Find all works in the two datasets for authors who have at least once published on the subject of "Political science".
- Find all works in these two datasets for authors who have at least once published works which are categorised as "biography".

Regarding the penultimate workset, it is of particular note that this returns results across both datasets, since our EEBO-TCP import did not contain genre or topic information; this association must be entirely inferred from the semantic links via the technology described below.

Implementation

Metadata requirements and ontology selection

Building on Nurmikko-Fuller et al. (2015) and Jett et al. (2016) we surveyed the addressable resources and the schema expressivity of ontologies that could parameterise these classes of workset. We identified parsable information structures in the EEBO-TCP TEI data, appropriate to the test worksets, and selected ontology terms to encode this EEBO-TCP metadata, ensuring compatibility (or at least, for our purposes, comparison) with RDF from the HathiTrust. The resultant ontology collection - the EEBO Ontology, or EEBOO - includes selections from MODS, Bibframe, and PROV, along with custom elements encoding additional structures (e.g. dates).

Creating EEBOO RDF and alignment with HTDL

Python scripts manipulated TEI P5 XML, then the Karma Data Integration Tool mapped EEBO-TCP data structures into the EEBOO ontology. Particular attention was paid to dates encoded within strings, an example of rich semi-structured data that can be extracted into structured RDF. Links to author records in VIAF and the Library of Congress (LoC), and multimedia pages in the HTDL and 'JISC Digital Books' website, were generated and added. Finally, author names were aligned between the EEBOO and HTDL triples using a reconfiguration of the SALT tool (Weigl et al. 2016).

24,926 EEBO-TCP Phase 1 records were processed, with 22 distinct types of information in the headers, including 6 different ID types and 3 types of date (publication date of historical work, author associated historical date(s), XML publication/editing dates). EEBOO incorporates 7 of these datatypes, and extends into subcategories for author names and date types. EEBOO contains 713 unique places, 6,489 unique expressions of Person of which 3,588 have VIAF and LoC IDs.

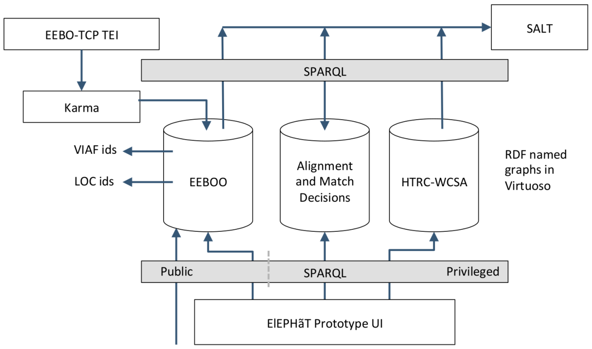
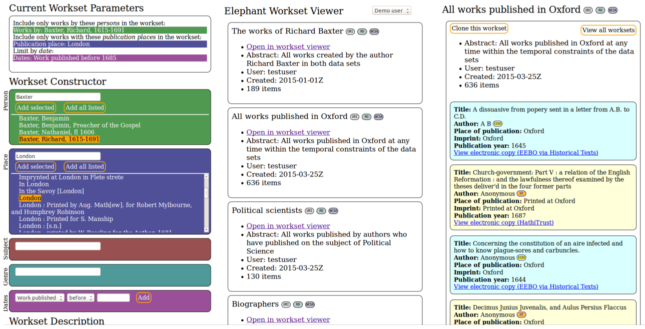


Figure 2. Architecture providing cross-corpus worksets for early English print

Workset construction and viewing

A Virtuoso triplestore (see also, the Virtuoso Github repository) stores the RDF data (totalling 1,137,502 triples) and provides a SPARQL query interface. Figure 2 shows the overall system architecture. The workset constructor user interface (Figure 3) allows the user to select parameters in a web interface which are, in the background, assembled into SPARQL queries used to create a workset. The interface automatically populates valid attributes that are themselves retrieved from the triplestore, using ontological terms having equivalent meaning across datasets. In combination, the generated triples and SPARQL queries are fully sufficient for expressing the motivating workset definitions described earlier.

The workset viewer (also Figure 3) then retrieves RDF workset contents, record metadata, data links, and multimedia links (to the Historic Books collection or the HTDL). Both web applications are written in Python, using the Flask framework, and both rely on the semantic information encoded in RDF and queried using SPARQL.



Current Workset Parameters

Include only works by these *persons* in the workset:
Works by: Baxter, Richard, 1615-1691
Include only works with these *publication places* in the workset:
Publication place: London
Limit by *date*:
Dates: Work published before 1685

Baxter, Benjamin
Baxter, Benjamin, Preacher of the Gospel
Baxter, Nathaniel, fl 1606
Baxter, Richard, 1615-1691

Imprynted at London in Flete strete
In London
In the Savoy [London]
London
London : Printed by Aug. Math[ew]. for Robert Mylbourne, and Humphrey Robinson
London : Printed for S. Manship
London : [s.n.]

Elephant Workset Viewer

The works of Richard Baxter

- Open in workset viewer
- Abstract: All works created by the author Richard Baxter in both data sets
- User: testuser
- Created: 2015-01-01Z
- 189 items

All works published in Oxford

- Open in workset viewer
- Abstract: All works published in Oxford at any time within the temporal constraints of the data sets
- User: testuser
- Created: 2015-03-25Z
- 636 items

Political scientists

- Open in workset viewer
- Abstract: All works published by authors who have published on the subject of Political Science
- User: testuser
- Created: 2015-03-25Z
- 130 items

Biographers

- Open in workset viewer

All works published in Oxford

- Abstract: All works published in Oxford at any time within the temporal constraints of the data sets
- User: testuser
- Created: 2015-03-25Z
- 636 items

Title: A dissuasive from popery sent in a letter from A.B. to C.D.
Author: A B
Place of publication: Oxford
Imprint: Oxford
Publication year: 1645
View electronic copy (EEBO via Historical Texts)

Title: Church-government: Part V : a relation of the English Reformation : and the lawfulness thereof examined by the theses deliver'd in the four former parts
Author: Anonymous
Place of publication: Printed at Oxford
Imprint: Printed at Oxford
Publication year: 1687
View electronic copy (HathiTrust)

Title: Concerning the constitution of an aire infected and how to know plague-sores and carbuncles.
Author: Anonymous
Place of publication: Oxford
Imprint: Oxford
Publication year: 1644
View electronic copy (EEBO via Historical Texts)

Title: Decimus Junius Juvenalis, and Aulus Persius Flaccus
Author: Anonymous
Place of publication: Oxford

Figure 3. Prototype workset constructor and viewer (example worksets shown)

Conclusion and future work

We have demonstrated the general feasibility of cross-corpus worksets in bringing together HathiTrust content with specialised collections through a specific implementation for early English printed books linking the HathiTrust to EEBO-TCP. Using Linked Data, we see that metadata can be extended in a piecemeal or iterative fashion, potentially moving beyond traditional bibliographic metadata to include semantic structures emerging from scholarly investigation of the worksets themselves; and in doing so support academic motivations and requirements for workset creation.

Acknowledgements

We are grateful to our colleague Pip Willcox for her valuable input and organisation of scholars' workshop, and Jacob Jett for his workset ontology. This work was supported by the Andrew W. Mellon Foundation through the Workset Creation for Scholarly Analysis project award.

Bibliography

Gibbs, F., Owens, T. (2012). Building better digital humanities tools: Toward broader audiences and user-centered designs. Digital Humanities Quarterly 6(2). Accessible via: http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html

Jett, J. (2015). Modeling worksets in the HathiTrust Research Center: CIRSS Technical Report WCSA0715. University of Illinois at Urbana-Champaign. Available via: http://hdl.handle.net/2142/78149

Jett, J., Nurmikko-Fuller, T., Cole, T.W., Page, K.R., Downie, J.S. (2016). Enhancing scholarly use of digital libraries: A comparative survey and review of bibliographic metadata ontologies. IEEE/ACM Joint Conference on Digital Libraries (JCDL) pp. 35-44, 2016.

Kambatla, K., Kollias, G., Kumar, V., Grama, A. (2014). Trends in big data analysis. Journal of Parallel & Distributed Computing 74(7), pp 2561-2573.

Nurmikko-Fuller, T., Page, K., Willcox, P., Jett, J. Maden, C., Cole, T., Fallaw, C., Senseney, M., Downie, J.S. (2015). Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications. ACM/IEEE Joint Conference on Digital Libraries (JCDL) p. 169-172, 2015.

Weigl, D. M., Lewis, D. L., Crawford, T., Knopke, I., Page, K. R. (2017, in press). On providing semantic alignment and unified access to music-library metadata. *International Journal on Digital Libraries.* Springer.

Contextual interpretation of digital music notation

Kevin Page
kevin.page@oerc.ox.ac.uk
University of Oxford, United Kingdom

David Lewis
david.lewis@oerc.ox.ac.uk
University of Oxford, United Kingdom

David Weigl
david.weigl@oerc.ox.ac.uk
University of Oxford, United Kingdom

Research Objects and scholarship in the digital age

As scientific research practice has grown to include ever greater quantities of data, larger collaborations, and distributed methods, Research Objects (Bechhofer et al, 2013) have been introduced as a means to gather together the context surrounding an investigation and to support its future validation, understanding, and re-use. In many cases these Research Objects build upon the methods, output, and provenance already captured and encoded by workflow systems -- the digital environments in which the science is conducted.

More recently there have been proposals for the use of Research Objects within the digital humanities and musicology (Dreyfus and Rindfleisch 2014; De Roure et al. 2016). Digital editions and annotations of digitally encoded works can be viewed as manifestations of workflows deployed in musicological scholarship, raising the question of how e.g. editorial annotations in a digital score should reference other digital items within the Research Object and vice versa.

For example, a study of the Mariinsky Opera's rendition of Wagner's *Ring* cycle in November 2014 produced a multimedia dataset including audio, annotations made by a musicologist on a short score before the performance (annotations which could be identified a priori from the score such as dynamics, appearances of a leitmotif, etc.), further annotations captured digitally by the musicologist during the live performance (typically staging and interpretive commentary), and free-form text from a digital pen (Page et al. 2016). If the score in use had been a digital edition encoded in MEI (the XML-based Music Encoding Initiative, as reviewed by Crawford and Lewis 2016) how might we reference the musicologist's annotations? Or the other media objects captured and the metadata describing them; existing Linked Data references to Wagner and leitmotifs; and to earlier surveys and studies made by the musicologist on leitmotif interpretation?

Linked notation in support of musicology

Notation examples are a vital part of analytical essays in musicology, helping to illustrate analytical observations and justify hypotheses, arguments and conclusions. They can be excerpts from a score, or custom-made notations which add annotations or comments to the original notation.

Furthermore, the presentation of multiple analogous musical extracts for comparison is often required to support a musicological narrative. *Paradigmatic analysis*, for example, involves passages of music placed one above another such that analogous elements are directly juxtaposed, with gaps left as necessary to ensure that vertical alignment is preserved. Stacked presentation of different scores or different parts of the same score have been used for well over a century, but they quickly become unwieldy and hard to interpret, especially as the number of extracts increases. What is not available in such paper-based approaches is the interactivity that can make complex comparisons between many extracts practical by turning a static presentation into an iterative exploration of digital materials.

In this paper we consider the example of a digital companion presenting the contents of a Research Object studying the interpretation of leitmotif examples from Wagner's compositions, specifically the *Ring* cycle, as they are presented in numerous historical introductions, opera guides, leitmotivic threads and leitmotif lists included in libretto editions and piano scores. The study of the incidences of these particular leitmotif identifications consists of both the gathering of source material and its digitisation and cataloguing, and a musicological study of the potential relationships, influence and evolution between leitmotif interpretations. To enable the extension and repurposing of the identified leitmotif relationships they are structured using an ontology.

Notation examples in leitmotif guides are usually abstractions drawn from a piano score. When reporting findings from this research it is desirable to present and relate the scholarly arguments back to the musicological context within which they are made: from the score excerpts in the source material; and to MEI encodings that illustrate and encode both the examples from which they are drawn and to complete (piano) scores of the overall operas.

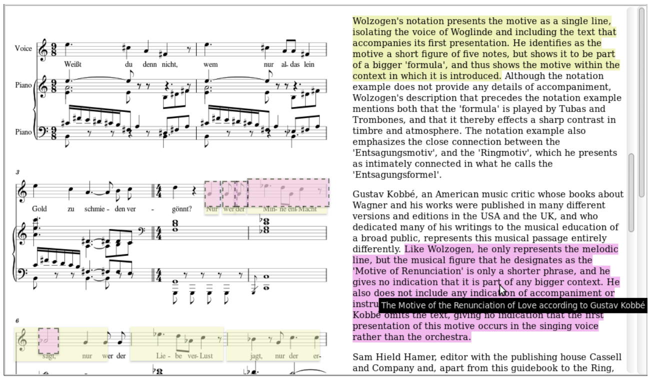
This enables matching and linking of the examples as they are described in the scholarly text, via semantic hyperlinks, to and from the score, including exact matches and variants, illustrating interpretations, and situating the examples back in context. Encoding interpretations in the form of notation examples as variant readings of a certain passage could thereby chart the ‘understanding’ of the work as a history of its variants.

(For example comparing: Richard Wagner, Die Walküre, piano score by Felix Mottl, Leipzig,

Peters, 1914, p.165; Hans von Wolzogen, Thematischer Leitfaden durch die Musik zu Richard Wagners Festspiel Der Ring des Nibelungen, 2nd ed., Leipzig, Schloemp, 1876, p.58, ‘Schicksalsmotiv’; Gustav Kobbé, Wagner's Music Dramas Analyzed With the Leading Motives, New York: Schirmer, 1923, p. 57, ‘Motive of Fate’; George Dunning Gribble, The Master Works of Richard Wagner, London, Everett, 1913, p. 289, ‘Fate Motif’.)

Introducing MELD: Music Encoding and Linked Data

To realise the digital notation companion we have developed the MELD framework (Music Encoding and Linked Data). MELD enables the interactive presentation of multimedia contents of the Research Object, such as the images, text, audio, and MEI encoded music notation described in the previous section. These can be explored contextually alongside each other through the use of semantic links, encoded using RDF, which describe the musicological relationships between the resources (and elements within them). In contrast to earlier technologies which have typically aligned resources against a timeline (e.g. in milliseconds, or using MIDI), MELD expresses relationships anchored to musically meaningful items scoped using MEI. Figure 1 shows a screenshot of MELD displaying text and notation, highlighting leitmotifs as identified in different historical guides.


Voice
Piano
Weißt du denn nicht, wem nur al- das lein
Gold zu schmie - den ver - gönnt?
nur wer der Lie - be ver-Lust

Wolzogen's notation presents the motive as a single line, isolating the voice of Woglinde and including the text that accompanies its first presentation. He identifies as the motive a short figure of five notes, but shows it to be part of a bigger 'formula', and thus shows the motive within the context in which it is introduced. Although the notation example does not provide any details of accompaniment, Wolzogen's description that precedes the notation example mentions both that the 'formula' is played by Tubas and Trombones, and that it thereby effects a sharp contrast in timbre and atmosphere. The notation example also emphasizes the close connection between the 'Entsagungsmotiv', and the 'Ringmotiv', which he presents as intimately connected in what he calls the 'Entsagungsformel'.

Gustav Kobbé, an American music critic whose books about Wagner and his works were published in many different versions and editions in the USA and the UK, and who dedicated many of his writings to the musical education of a broad public, represents this musical passage entirely differently. Like Wolzogen, he only represents the melodic line, but the musical figure that he designates as the 'Motive of Renunciation' is only a shorter phrase, and he gives no indication that it is part of any bigger context. He also does not include any indication of accompaniment or instru

The Motive of the Renunciation of Love according to Gustav Kobbé

Kobbé omits the text, giving no indication that the first presentation of this motive occurs in the singing voice rather than the orchestra.

Sam Hield Hamer, editor with the publishing house Cassell and Company and, apart from this guidebook to the Ring,

Figure 1. MELD displaying contextualised text and music notation.

To render our music notation (encoded using MEI) we use Verovio (Pugin et al. 2013), an open-source MEI renderer that produces beautiful SVG renditions of the score. In addition, Verovio provides an architecture in which identifiers (in other words, anchors for our relational Linked Data expressions in the MEI XML) are persisted through to the rendering (in SVG) which can be connected to identifiers in our contextual information (in RDF). When rendered (and re-rendered) for the user in our web based application interface, the browser uses these identifiers to generate interface elements and undertake actions that combine information from the MEI and the Linked Data.

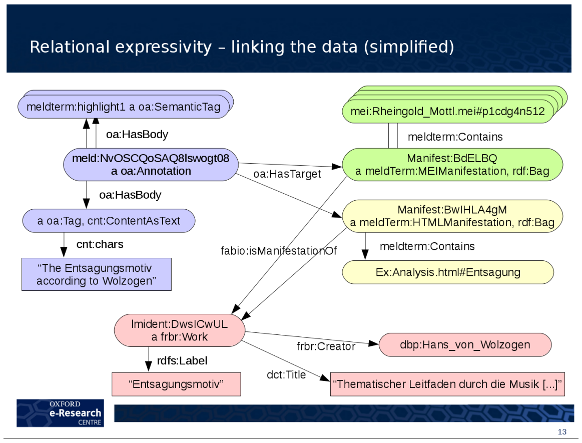

Relational expressivity – linking the data (simplified)

Figure 2. Musicological relationships, encoded using Open Annotations, within the Research Object (simplified).

Within the Research Object, we treat the XML IDs of elements within the MEI resource as fragment identifiers, so URIs can be straightforwardly generated for each notation element of interest. We employ the Web Annotation Data Model (Sanderson et al. 2017), using these URIs as targets of annotations representing each musicological marking. Corresponding annotation bodies are associated with semantic tags defined to encode the different musicological interpretations, which are in turn the annotation bodies of a top-level annotation targeting the URI of the files currently being viewed, including the music encoding (MEI) and scholarly interpretation (HTML). A simplified example of such relationships is shown in Figure 2.

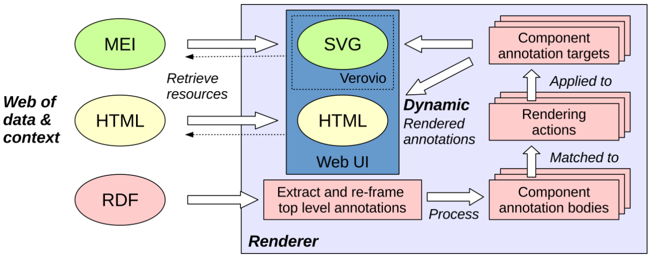


Figure 3. The MELD framework (shading corresponds to that in Figure 2).

The MELD client then uses HTML/CSS and JavaScript, served by a simple web service implemented with Python Flask, and illustrated in Figure 3. The procedure driving the rendering and user interaction is illustrated in Figure 1. The client processes a framed (see the explanation of framing) JSON-LD representation of the RDF graph instantiating the data model. It then performs an HTTP GET call to acquire the MEI resource targeted by the top-level annotation, and renders the corresponding musical score to SVG using Verovio. User interactions are captured using HTML divs drawn as bounding boxes over portions of the SVG corresponding to MEI elements of interest; this is simplified by

Verovio's retention of MEI identifiers in the produced SVG output.

Acknowledgements

This work has been supported by the UK Arts and Humanities Research Council grant AH/L006820/1 'Transforming Musicology', and EPSRC grant EP/L019981/1 'Fusing Audio and Semantic Technologies for Intelligent Music Production and Consumption'. We thank all our colleagues on these projects who have supported and encouraged this work, particularly Carolin Rindfleisch for her complementary work on musicological ontologies, and Laurence Dreyfus for wider inspiration and motivation in developing these technologies.

Bibliography

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I. and Gamble, M. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), pp.599-611.

Crawford, T., & Lewis, R. (2016). Review: Music Encoding Initiative. *Journal of the American Musicological Society*, *69*(1), 273-285.

De Roure, D., Klyne, G., Page, K.R., Pybus, J., Weigl, D.M., Wilcoxson, M., Willcox, P. (2016). Plans and Performances: Parallels in the Production of Science and Music. *Proceedings of the 2016 IEEE 12th International Conference on e-Science*. IEEE, pp. 185-192.

Dreyfus, L., & Rindfleisch, C. (2014). Using Digital Libraries in the Research of the Reception and Interpretation of Richard Wagner's Leitmotifs. *Proceedings of the 1st International Workshop on Digital Libraries for Musicology.* ACM, pp. 1-3

Page, K., Nurmikko-Fuller, T., Rindfleisch, C., Weigl, D. (2016). Digital Annotation Tooling for Opera Performance Studies. *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 306-309.

Pugin, L., Zitellini, R., & Roland, P. (2014). Verovio: A Library for Engraving MEI Music Notation into SVG. *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pp. 107-112.

Sanderson, R., Ciccarese, P., Young, B. (2017). Web Annotation Data Model. W3C recommendation.

Twitter as a source for the history of the present : the 2015 Greek referendum as a case study

Sofia Papastamkou
sofia.papastamkou@meshs.fr
Maison Européenne des Sciences de l'Homme et de la Société, France

Introduction

Twitter data have been growingly used as a source for scholarly studies in various disciplines in recent years (Williams et al., 2013). The value of such born-digital data as primary source materials for future researches in history is already being acknowledged (Webster, 2015, Steinhauer, 2015). But, at least for now, historians seem rather reluctant to make use of them, although some recent works deal with the perception and memory of the past on Twitter (Clavert, 2016, Turgeon, 2014) or propose both documentation and analysis of present time events (Ruest and Milligan, 2016).

One possible reason of this reluctance could be the attachment of historians to traditional archival collections, e.g. those organized by professional archivists. But the creation of archives for social network sites data is not yet systematic and still in the beginning. As for the global Twitter archive of the Library of Congress, it is unknown when it will be functional (Zimmer, 2015). As it is possible for one to retrieve Twitter datasets, a second reason of this reluctance could be the need for acquaintance with basic methods and tools for gathering, understanding and analyzing born-digital data. However, not all historians are trained to digital humanities and quantitative methods that provide for such skills. Last but not least, the main reason could be the relation historians have with time. It has always been difficult to define the moving frontier between the present time and the recent past in contemporary history (Bédarida, 2003). As instant ephemera data that belong in the very present time, Twitter data precisely underline the difficulty for historians to define their own territory in these temporalities.

Nonetheless, for historians concerned with contemporary historical events – historical in the sense of a conjuncture that reveals a before and an after (Le Goff, 1999) – Twitter provides an original documentation. This documentation is generated in real time; organized around folksonomies – the hashtags – that reveal a direct perception from below; but also in close relation with media coverage. Since the creation of Twitter, a series of hashtagged global events (#IranElection, #15M, #Occupy, the 2011 Arab revolutions under various hashtags, the 2015 terrorist attacks in Paris...) update the concept of the monster-event (Nora, 1972) in that they are produced, lived, transmitted and shared in real-time around the globe – or at least in its connected parts. In spite of the known biases, mainly the fact that it is mainly used by relatively young and highly educated adults (Pew Research, 2016), Twitter offers an original kind of non-institutional primary sources (tweets) that can be complementary to the traditional ones the historians use.

This paper is a tentative to document and to provide a first analysis, based on Twitter primary sources, of the Greek referendum of 2015. This event has already obtained a distinctive status in the "before" and the "after" the current crisis marks in Greece's post-transition to democracy history (after 1974) (Avgheridis et al., 2015), although time and future historians will definitely tell. The paper considers the transnational phase of this event, which followed the Greek vote in favour of the "no" to further austerity measures, included negotiations in the instances of the EU and ended with the agreement of the Greek government to conclude a third harsh austerity programme. Our main research hypothesis is that the imbrication of different hashtags reveals different temporalities that allow researchers to construct regimes of historicity of an event.

Event background

In the aftermath of the 2008 financial crisis, Greece, an EU and Eurozone member, began going through a severe debt crisis that revealed the structural weaknesses of the European monetary union and soon expanded to other weak members (Portugal, Ireland, Cyprus and Spain). Since 2010, the crisis has been managed through the setup of a European financial assistance mechanism in exchange for national programmes of structural reforms and budgetary cuts. Two such programmes were applied to Greece in 2010 and in 2012, plus a debt restructuring, that were monitored by the European Commission, the European Central Bank and the International Monetary Fund (Papaconstantinou, 2016; Zettelmeyer, 2013). The ongoing crisis provoked profound social and political transformations in the country that brought to power a coalition government led by the radical Left party of Syriza in January 2015. Syriza won the election with the promise to put an end to austerity politics. The party emerged in the context of the post-2008 crisis that shook the countries of Southern Europe and the 2011 Indignant movements, just like Podemos in Spain. Thus, the referendum of July 2015 was far from being significant only in the context of the Greek crisis as an effort of the new government to ameliorate the terms of the Greek programmes, as it put at stake different visions for the EU and its crisis management politics.

Data collection and analysis: method and tools

Tweets using the hashtag #greferendum were collected with NodeXL, an add-in to MS Excel (Smith et al., 2009). The collect was setup once daily from July, 6 to July, 16 2015. The size of the gathered sample was determined by the capacities of the tool, that can collect a maximum of around 20,000 tweets at once. A total of 204,714 tweets were collected of which 139,945 are retweets (68,36 %), 8, 686 responses (4,24 %), 56,086 mere tweets (27,39 %). Minor collects were also launched for other related hashtags (mainly #thisisacoup). Hashtag data were treated with OpenRefine and further explored with R software.

Statistical analysis of textual data (tweets) was made with TXM-Textometry software (Heiden, 2010). The corresponding dataset had been previously encoded following the TEI P5/XML standard with use of the OxGarage service.

Social network analysis and visualizations were made with Gephi software (Bastian et al., 2009).

Data analysis: first findings

Hashtags

The first part of the research focused on reading the hashtags of the dataset. The hashtag #greferendum was used with a variety of hashtags, a total of some 12,000 words (all languages and variants included). A first study focused on the hashtags with a frequency over 99, which gave a total of 158 words. After an elementary typology was established, it was possible to distinguish: geographic names, names of persons, institutions, common names, neologisms that came out of contractions (such as "greferendum"), short phrases that had the function of commentary. The use of hashtags varied between tag and commentary, or included both functions at once (Bruns and Burgess, 2011). The big majority of hashtags are in English (112 out of 158). However, in the thirty most frequent hashtags of the dataset, it is possible to find words in Spanish, Italian, French, and German. By consequence, the linguistic communities that participated in the global interactions were the ones that were the most concerned by the crisis. As for the Greek language, it is not entirely absent as such, but it is mainly present in its greeklish form: Greek words used as hashtags but written in Latin alphabet.

A close reading of the thirty most frequent hashtags with parallel consideration of the associations of words (coocurrencies) shows the tractations that followed the Greek referendum were basically perceived as an intergovernmental affair with the EU actors occupying a secondary position.

An interesting case is the emergence of the hashtag #thisisacoup as an act of solidarity of Spanish militants of *Barcelona en Comú* towards the Greek government during the Eurogroup and the Euro Summit negotiations (12-13 July). The corresponding dataset is more oriented to the expression of personal opinion than the dissemination of information with hashtags in the form of phrases that function more as commentary than tags (such as #yovoycongrecia).

Domains

The most tweeted domains were twitter.com (7,352 tweets) and theguardian.com (7,217 tweets). In general, it is possible to distinguish two main tendencies. First, the dissemination of information in social media (Twitter, YouTube, Instagram, Facebook). Second, the dissemination of authoritative information (international media, specialized independent blogs, personal blogs).

Communities detection

The network of the #greferendum corpus is composed by 103,733 nodes and 204,713 relations. After nodes with a degree higher than 10 were isolated (around 4% of the total), 326 communities were detected with Gephi (Louvain method). These communities need to be further explored, however the first findings for the most important of them show that affinities developed around sources of information (media), political and/or intellectual personalities, professional communities, and also linguistic communities.

Conclusion

The network and the detected communities seem to have been structured around the dissemination of information but also political affinities and/or militantism. However, further exploration is necessary in order to better understand the network structure.

A quantitative analysis of the tweets, with emphasis on the associations between the hashtags, indicate coexistence of different temporalities within the temporality of the 2015 Greek referendum that are principally related to the Eurozone crisis, the associated national sub-crisis, and post-2008 anti-austerity movements. In this sense, Twitter primary sources offer insights from a transnational scale.

Bibliography

Avgheridis, M., Gazi, E. and Kornetis, K. (eds) (2015). *Μεταπολίτευση. Η Ελλάδα στο μεταίχμιο δύο αιώνων* [Metapolitefsi. Greece Between Two Centuries]. Athens: Themelio, pp. 15-18 and 335-66.

Bastian, M., Heymann, S., Jacomy, M. (2009). “Gephi: An Open Source Software for Exploring and Manipulating Networks.” *International AAAI Conference on Weblogs and Social Media*. San Jose, California: Association for the Advancement of Artificial Intelligence, pp. 361-62

Bédarida, F. (2003). *Histoire, critique et responsabilité*. Brussels: Complexe, p. 64

Bruns, A. and Burgess, J. (2011). “The Use of Twitter Hashtags in the Formation of Ad Hoc Publics”. *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference*. Reykjavik: University of Iceland. Available at: http://eprints.qut.edu.au/46515/

Clavert, F. (2016). “#ww1. The Great War on Twitter.” *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 461-62. Available at: http://dh2016.adho.org/abstracts/378

Hanneman, R. A. and Riddle, M. (2005). Introduction to social network methods. Riverside, California: University of California, 2005

Heiden, S. (2010). “The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme.” *24th Pacific Asia Conference on Language, Information and Computation*. Sendai: Institute for Digital Enhancement of Cognitive Development, Waseda University, pp.389-398. Available at: https://halshs.archives-ouvertes.fr/halshs-00549764/document

Le Goff, J. (1999). “Les « retours » dans l'historiographie française actuelle”. *Les Cahiers du Centre de Recherches Historiques*, 22. Available at: http://ccrh.revues.org/2322 ; DOI:10.4000/ccrh.2322

Nora, P. (1972). "L'événement monstre". *Communications*, 18: 162-72

Papaconstantinou, G. (2016). Game Over: The Inside Story of the Greek Crisis. Middleton, Delaware: CreateSpace

Pew Research Center (2016). “Social Media Update 2016”. Available at: http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Ruest, N. and Milligan, I. (2016). “An Open Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter.” *Code4Lib Journal*, 32. Available at http://journal.code4lib.org/articles/11358

Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M., Barash, V., Dunne, C., Capone, T., Perer, A. and Gleave, E. (2009). “Analyzing (Social Media) Networks with NodeXL.” *Proceedings of the Fourth International Conference on Communities and Technologies*. New York: ACM, pp. 255-64

Steinhauer, J. (2015). “Preserving Social Media for Future Historians.” *Insights. Scholarly Work at the John W. Kluge Center*. Available at: https://blogs.loc.gov/kluge/2015/07/preserving-social-media-for-future-historians/

Turgeon, A. (2014). “Comment travailler la mémoire sur Twitter”. *Études canadiennes / Canadian Studies*, 76. Available at: http://eccs.revues.org/216

Webster, P. (2015), “Will Historians of the Future Be Able to Study Twitter ?”. *Webstory, Peter Webster’s Blog*. Available at: https://peterwebster.me/2015/03/06/future-historians-and-twitter/

Williams, Sh.-A., Terras, M. and Warwick, C. (2013). “What people study when they study Twitter: Classifying Twitter related academic papers.” *Journal of Documentation*, 69 (3): 384-410

Zettelmeyer, J. , Trebesch, Ch. and Gulati, M. (2013) “The Greek debt restructuring: an autopsy”. *Economic Policy*, 28 (75): 513-63.

Zimmer, M. (2015). “The Twitter Archive at the Library of Congress: Challenges for information practice and information policy." First Monday 20 (7). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/5619/4653

L’archéologie du paysage sonore: l’Histoire à portée de nos oreilles

Mylène Pardoen
mylene.pardoen@wanadoo.fr
Institut des Sciences de l'Homme de Lyon, CNRS, France

L’archéologie du paysage sonore est un concept très récent. Elle tend à chercher/pister les traces sonores passées et disparues, tant dans les écrits que dans les visuels et à les restituer. Elle propose un modèle virtuel qui n’incarne pas une vérité scientifique figée mais une proposition représentant l’état de la science à l’instant de sa création et qui est voué à évoluer. C’est donc la restitution d’une possible réalité d’un quotidien passé qui s’appuie sur des critères scientifiquement valides.

L’archéologie du paysage sonore innove à différents niveaux et ses spécificités portent sur :

- L'inter et la transdisciplinarité (association étroite entre les Sciences Humaines et Sociales et les Sciences de l'ingénieur) ;
- La valorisation du patrimoine par le sensible;
- La construction autour du son et une élaboration concomitante visuel/son.
- Elle comporte également des innovations en matière d'édition scientifique :
- nouveaux procédés éditoriaux de type hétérographique ;
- sonorisation de nouveaux médias impliquant la réalité augmentée.

L'archéologie du paysage sonore est une réponse possible pour les musées, les maquettes à destination ludo-pédagogique et culturelle ainsi que tout support à même destination. Son caractère scientifique la distingue du sound design (plutôt axé composition artistique et bruitage (pour le cinéma pour le théâtre).

Et, dans un premier temps, c'est pour satisfaire à la demande spécifique des musées en matière de « mise en sons » que les premières réflexions en ce sens sont menées, après avoir fait le constat que les réponses apportées ne leur donnaient pas totalement satisfaction.

Cette nouvelle branche des sciences humaines et sociales prend ses racines dans les études de R. Murray Schafer et son concept d'écologie du sonore (Le paysage sonore – le monde comme musique, 1977, réédition 2010) et dans celles sur le sensible menées par A. Corbin (Les cloches de la terre, 1994, et plus récemment Histoire du silence, 2016). Cette prise de conscience des spécificités de notre environnement qui forment un paysage sonore est à compléter par une petite étude de J. P. Gutton (Bruits et sons dans notre histoire – 2000). Mais toutes ces études restent à l'état théorique. Or, de nos jours, les technologies issues du multimédia nous offrent la possibilité de passer à l'étape expérimentale et de proposer des restitutions respectueuses de l'Histoire et de ses cadres scientifiques.

S'appuyer sur le sensible, et notamment le sonore, pour valoriser le patrimoine permet de répondre à une demande souvent mal définie des conservateurs de musée ou des services d'archives qui désirent associer un volet multimédia lors de leurs expositions à des fins de rendre moins austère la présentation de leurs riches collections. En effet, à une époque où le multimédia modifie chaque activité de découverte, agissant notamment sur la perception, il devient nécessaire d'intégrer cette dimension du sensible lors des présentations touchant au patrimonial.

Que ce soit dans le cadre d'une restitution de paysages urbains, d'ambiances d'intérieur, ou autres, les problématiques restent identiques. La restitution ou création de paysages sonores historiques pose, d'entrée, des questions fondamentales : Peut-on « entendre » ce passé ? Comment le restituer, le faire entendre ? Quelles sont les difficultés, les limites d'une telle restitution ? Quels sont les matériaux ? Où s'arrête l'acte du créateur-artiste-compositeur, celui du chercheur ? Comment dépasser le concept d'habillage sonore (sound design) et aboutir la restitution d'un véritable paysage historique ? Comment intégrer ces notions relevant du sensible, les porter au public ? Pour quel public ?

La demande muséographique place le concepteur aux confins de la composition, de la musicologie et de l'histoire. Un « simple » habillage sonore (sound design) ne peut, en effet, satisfaire cette demande très spécifique qui entre, également, dans les cadres de la sauvegarde du patrimoine immatériel.

En effet, pour le musicologue, le compositeur ou le metteur en sons, ce travail est inhabituel. Les matériaux ordinaires sont quasi inexistants : pas d'enregistrements disponibles pour les années antérieures à 1870. Entendre le passé implique donc qu'il faut en chercher les traces dans l'ensemble du corpus littéraire (livres, journaux, etc.) ou graphiques. Puis, tenter de le pister dans le temps d'aujourd'hui afin d'en faire une captation. Puis le restituer de manière à le rendre compréhensible et en neutralisant le discours afin de faire acte de témoignage (une narration d'historien et non une œuvre de compositeur, où l'acte émotionnel est important et conduit le discours musical).

Pour le visiteur, l'ambiance sonore ainsi (re)créée doit guider son imaginaire pour l'aider à mieux percevoir, recontextualiser des situations historiques, tout en préservant la réalité historique et le laisser maître de ses ressentis et émotions personnelles : trouver le juste milieu, l'équilibre dans la sollicitation de cet imaginaire pour mieux servir le patrimoine et le mettre en valeur.

Entre le constat ou l'émission d'un besoin et son résultat, se situe le travail de l'archéologue du paysage sonore. Dès les prémices, il faut recueillir un certains nombre d'informations, notamment techniques, auprès des conservateurs afin de cerner au mieux la forme que peut ou doit prendre le rendu sonore. L'expérience nous a montré que de nombreuses contraintes techniques étaient présentes, notamment des contraintes matérielles et logicielles. Parmi celles-ci, l'obsolescence présente un réel frein pour les musées car elle représente un coût financier important. Il devient donc évident de mettre au service des musées, des historiens ou des archivistes une représentation audio qui respecte leurs propres contraintes techniques.

Puis viennent les deux phases de « récolte ». La première nécessite le moissonnage de sources hétérogènes (textuelles diverses et visuelles). Initialement artisanale (faite manuellement), actuellement, nous nous orientons vers la mise au point d'outils d'extraction multisources s'appuyant sur le Text-Mining et la fouille d'image de document (Data Mining), ainsi que l'élaboration ontologies grâce à l'appui de nos collègues des sciences de l'ingénieur. D'autre part, des protocoles seront élaborés pour la phase captation, actuellement plutôt expérimentale. La collecte sonore se fait par captation de sons et d'ambiances réelles – preuve des traces de ce passé dans notre présent.

La dernière phase reste, pour l'instant manuelle. La restitution des ambiances sonores doit « caler » au plus proche du support et de la demande et peut prendre des formes différentes, selon que l'on travaille pour un espace, un support vidéo ou un support maquette. Cette partie du travail prend donc des formes différentes et leur mise en oeuvre empruntent à la fois au cinéma et au jeu vidéo – tout en visant un résultat différent, puisque la restitution se veut ne pas véhiculer d'émotions, mais uniquement susciter des ressentis.

La présentation s'articulera sur les différentes phases d'élaboration. Les présentations sonores seront issues de mes travaux :

- la maquette qui sert de matrice à la recherche (*Bretez II* – projet de restitution de Paris au XVIIIe siècle.). Ce projet relève de la recherche empirique et d'une méthodologie expérimentale ayant pour objectif la mise au point d'un modèle.
- Mes travaux sur les fresques sonores historiques, qui entrent dans le cadre de commandes pour les musées (musée archéologique de la crypte Notre-Dame/Carnavalet, Le Louvre et Versailles) permettent également de nourrir ma réflexion.

C'est l'ensemble de ces expériences qui me permettent d'élaborer mes trames et méthodologies.

Literary Exploration Machine: A New Tool for Distant Readers of Polish Literature

Maciej Piasecki
maciej.piasecki@pwr.edu.pl
Wrocław University of Science and Technology
Poland

Tomasz Walkowiak
tomasz.walkowiak@pwr.edu.pl
Wrocław University of Science and Technology
Poland

Maciej Maryl
maciej.maryl@ibl.waw.pl
Polish Academy of Sciences, Poland

Brief Summary

This paper presents an initial prototype of a web-based application for textual scholars. The goal of this project is to create a complex and stable research environment allowing scholars to upload the texts they are analysing and either explore with a suite of dedicated tools or transform them into another format (text, table, list). This latter functionality is especially important for research into Polish texts, because it allows for further processing with the tools built for the English language.

This application brings together the existing applications developed by CLARIN-PL and supplements them with new functionalities. The project is based on a close cooperation between IT professionals, linguists and literary scholars, which ensures that the tools will suit actual researchers' needs.

The main features of LEM include: lemmatization, part-of-speech tagging, text clustering, semantic text classification based on machine learning, and visualisation of its output, generating custom wordlists and lemmatized texts.

Challenge

Digital literary studies seem to be one of the most vividly developing strand of digital humanities. Different analytical systems were proposed, e.g. Mallet, PhiloLogic3 plus PhiloMine, but focused on selected techniques and mostly on English texts. Their language-processing capabilities are limited only to lemmatization and morphosyntactic tagging and they usually require from their users certain programming skills.

In order to address those challenges we have developed a prototype of a web-based system, called *Literary Exploration Machine* (LEM), which does not require installation and programming skills. LEM has a component-based architecture, remains open for expanding components, implements natural language processing on different levels and is planned to support several different paradigms of the text analysis.

Scheme of the system

Components

Word frequencies can be simply computed for English, but not for highly inflected languages such as Polish, which has more than 100 possible word forms of an adjective (however, almost-full sets of distinct forms exist only for some lemmas). In such languages, morphological forms have to be first mapped to *lemmas* by a morpho-syntactic tagger, e.g. WCRFT2 for Polish (Radziszewski, 2013). By applying different language tools, we can enrich texts with metadata revealing linguistic structures.

LEM expands WebSty - an open stylometric system, adopting the following features for text description: segmentation-based (lengths of documents, paragraphs and sentences), morphological (words, punctuations, pseudo-suffixes and lemmas), grammatical classes and categories (e.g. from the Polish National Corpus –see Przepiórkowski et al, 2012– tagset, Broda and Piasecki, 2013) and their n-grams.

This set has been additionally expanded in LEM with the following features, allowing for semantic analysis:

- semantic *Proper Name classes* – recognised by a Named Entity Recogniser Liner2 (Marcińczuk et al, 2013),

- temporal, spatial relation (Kocoń and Marcińczuk, 2015), and selected semantic binary relations (e.g. *owner of*) ,
- *lexical meanings* – synsets in plWordNet (the Polish wordnet); assigned to words and selected multiword expressions by Word Sense Disambiguation tool WoSeDon (Kędzia et al, 2015),
- generalised lexical meanings – meanings mapped to more general synsets, e.g. *an animal* instead of *a cheetah*,
- lexicographic domains from Wordnet.

Rich text description is a good basis for several processing paradigms that LEM is going to support, namely:

- *linguistic text preprocessing* - extraction of language data for further statistical analysis, i.e. computing frequencies as the initial feature values, e.g., of lemmas, tags, word senses, etc.,
- *topic modelling*,
- unsupervised *semantic text clustering* and analysis of characteristic features for clusters,
- supervised *semantic text classification* - trained on the manually annotated texts,
- stylometric analysis - performed with the help of the WebSty system.

Processing scheme

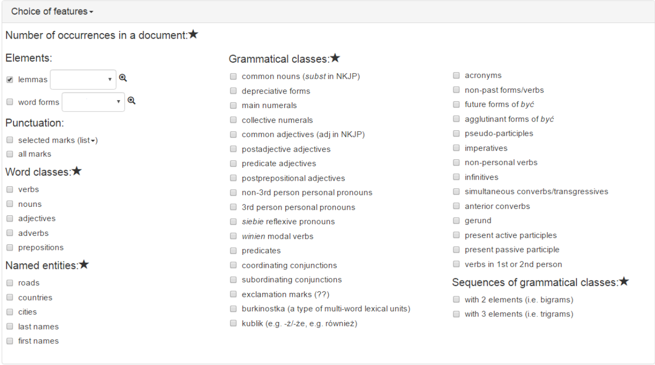
The processing paradigms share the following workflow:

- Uploading a corpus of documents together with metadata in CMDI format (Broeder et al, 2012) from the CLARIN infrastructure.
- Text extraction and cleaning.
- Choosing the features for the description of documents by users (see Fig. 1).
- Setting up the parameters for processing (users).
- Pre-processing texts with language tools.
- Calculating feature values for the pre-processed texts.
- Filtering and/or transforming the original feature values.
- Data mining.
- Presenting the results: visualisation or export of data.

To facilitate the upload, users are encouraged to deposit large text collections in the CLARI-PL dSpace repository. Users are advised to use public licences, but private research corpora can be also uploaded.

OCR-ed documents usually contain many language errors that should be corrected to some extent in the step 2. Moreover, metadata elements (e.g. page numbers, headers and footers) have to be separated during from the content and stored in a standalone annotation.

Users are not expected to have advanced knowledge of Natural Language Engineering or Data Mining. Thus, in Step 4, default settings of parameters will be provided. More advanced users will be able to tune the tool to their needs (see Fig. 1)



Choice of features ▾

Number of occurrences in a document:★

Elements:

☑ lemmas
□ word forms

Punctuation:

□ selected marks (list ▾)
□ all marks

Word classes:★

□ verbs
□ nouns
□ adjectives
□ adverbs
□ prepositions

Named entities:★

□ roads
□ countries
□ cities
□ last names
□ first names

Grammatical classes:★

□ common nouns (*subst* in NKJP)
□ depreciative forms
□ main numerals
□ collective numerals
□ common adjectives (adj in NKJP)
□ postadjective adjectives
□ predicate adjectives
□ postprepositional adjectives
□ non-3rd person personal pronouns
□ 3rd person personal pronouns
□ *siebie* reflexive pronouns
□ *winien* modal verbs
□ predicates
□ coordinating conjunctions
□ subordinating conjunctions
□ exclamation marks (??)
□ burkinostka (a type of multi-word lexical units)
□ kublik (e.g. -ż/-że, e.g. również)
□ acronyms
□ non-past forms/verbs
□ future forms of *być*
□ agglutinant forms of *być*
□ pseudo-participles
□ imperatives
□ non-personal verbs
□ infinitives
□ simultaneous converbs/transgressives
□ anterior converbs
□ gerund
□ present active participles
□ present passive participle
□ verbs in 1st or 2nd person

Sequences of grammatical classes:★

□ with 2 elements (i.e. bigrams)
□ with 3 elements (i.e. trigrams)

Figure 1. Web interface - a panel with a list of features

In Step 5 language tools are run. Each text is analysed by a part-of-speech tagger (e.g. WCRFT2) and next piped to a name entity recognizer (e.g. Liner2, Marcińczuk et al, 2013), temporal expression recognition, word sense recognition (WoSeDon, see Kędzia et al, 2015), etc.

Extraction of features encompasses counting frequencies, but also annotations matching patterns for every position in a document. In the case of wordnet-based features, meaning generalisation is done by iterating via wordnet structure.

A dedicated feature extraction module was built that is similar to Fextor (Broda et al, 2013) but much more efficient by supporting parallel processing. As a result of Step 6 every document is represented as vector of feature values and/or a sequence of language elements.

Filtering and transformation functions comes from the clustering packages or dedicated systems, e.g. SuperMatrix system (Broda and Piasecki, 2013).

Step 8 differentiates between the processing paradigms. Topic modelling, e.g. by Mallet, takes documents represented as lemma sequences. They can be also processed by corpus tools, e.g. for concordances and frequencies. Documents as feature vectors can be processed by clustering systems e.g. Cluto, or used in machine learning, e.g. Weka system.

Different processing paradigms provide varied perspectives on the data, e.g. topic modelling represents a document in terms of stochastic processes generating word occurrences from topic-related subsets in the text. Clustering reveals groups of documents based on content similarity. It is difficult to find a system that supports all paradigms.

In LEM, clustering is expanded with the extraction of features characteristic for the individual clusters. Several functions (from Weka, scikit-learn and SciPy packages), based on mathematical statistics, information theory and machine learning, are offered. The rankings of features are presented on the screen for interactive browsing and can be downloaded.

WebSty, based on elements of the same framework, can be applied to stylometric analysis.

Step 9, visualisation of clustering results (see Fig. 4), is based on Spectral Embedding (also known as Laplacian Eigenmaps). The 3D representation of the data (represented by similarity matrix) is calculated using a spectral decomposition of the graph Laplacian. Texts similar to each other are mapped close to each other in the low dimensional space, preserving local distances.

Use Case

The LEM prototype was developed by the team working with a particular textual corpus of 2553 Polish texts, published in *Teksty Drugie*, an academic journal dedicated to literary studies. The corpus consisted two parts: OCRd scans (1990-1998) and digital files (1999-2014). Given the aim of this paper (software presentation) and the shortage of space, we will treat the results only as examples of the method, without getting into too much detail.

The work on the prototype was divided into stages, conceived as a feedback loop for the developing team: on every stage a new service was added to application and the test run was performed. After the analysis of the result, the step was repeated or the team moved to the next phase.

Phase 1. Cleaning. The OCR-ed corpus has been cleaned (e.g. wordbreaks and headers were removed)

Phase 2. The corpus was lemmatized and parts of speech were tagged. Frequency lists were created what enabled the search for patterns in the textual output. For instance, Figure 2 shows the pattern of interest in particular Polish poets throughout 25 years, based on lemmatized mentions.

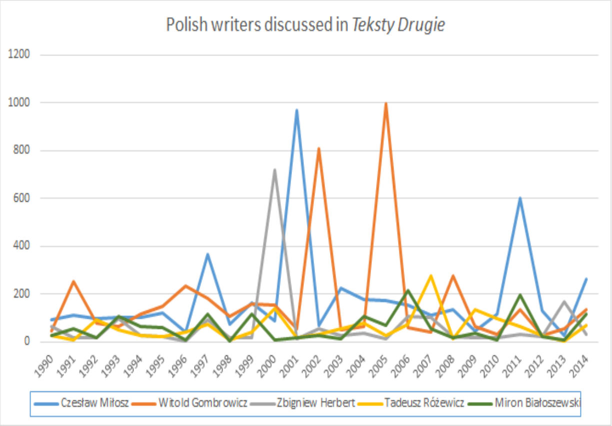


Figure 2. Pattern of interest in particular Polish writers in *Teksty Drugie* (1990-2014).

Phase 3. The analysis of the word frequencies revealed some problems with the word list, especially with numbers, years and city names, which were preserved in bibliographic references. A functionality of adopting a custom stopword list was employed. The exclusion of corpus-specific problematic words and general meaningless words (e.g. a, this, that, if) allowed for visualisation of the most frequent words in *Teksty Drugie* (Fig. 3)

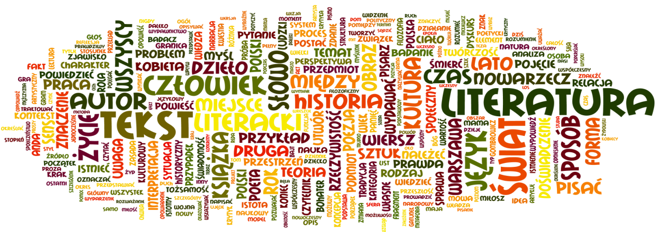


Figure 3. 300 most frequent words from *Teksty Drugie* (1990-2014) (meaningless words excluded) visualised with wordle.

Phase 4. The texts were then grouped into clusters of 20, 50 and 100 in a series of experiments. Each grouping revealed a bit different level of generalization about the texts. LEM, thanks to visualisation features (Fig. 4), allows for real-time exploration of deeper relationships between the texts.

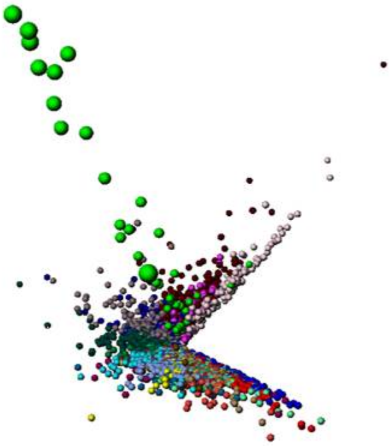


Figure 4. Visualisation of clustering results (weighting: MI-simple, similarity metric: ratio, number of clusters: 20, clustering method: agglomerative, visualization: the similarity matrix converted to distances and mapped to 3D by a spectral decomposition of the graph Laplacian - spectral embedding method).

By choosing the level of granularity (20, 50 or 100 clusters) we may analyse diverse patterns of discursive similarities between texts. Table 1 shows the differences in clustering of the same sample. The first option (20) shows the similarity between texts on a rather general level, that could be described as stylistic or genre similarity (e.g. formal vocabulary). Other options allow for more detailed exploration of general research approach (50) or particular topics analyzed in articles (100). Semantics of clusters is described by the identified characteristic features.

Number of clusters	100	50	20
Cluster size (mean)	25.33	50.66	56.65
Cluster size (median)	24	47	51,5
Smallest cluster size	13	25	2
Largest cluster size	51	91	96

Table 1. Differences between the clustering options (numbers reflect the quantity of texts assigned to particular cluster)

Researchers may explore all options and analyse the vocabulary responsible for classifying particular texts into a certain group by a virtue of being over- or under-represented in comparison to the entire sample.

The LEM is not a real time system. However, processing of the exemplar corpus (2553 documents from “Teksty Drugie”) takes less than 20 minutes. This is due to the use of a private cloud and proprietary message-oriented engine for processing texts. We plan to speed up the process, by running larger number of instances of language tools and by compressing results at each stage. Moreover, the user is able to start processing from any stage, so the processing time is shorter when the user plays with different settings.

Further Development

Currently LEM’s GUI is developed in cooperation with potential users, literary scholars working on various types of texts (fiction, journal articles, blog posts). That is also why we call this software “literary”, because further development will address the issues pertinent for literary theory, exceeding a purely linguistic perspective. Some literary-specific issues and functions will be expanded on the later stage of development, e.g. with adding language tools for Word Sense Disambiguation and partial analysis of the text structure, like anaphor resolution and discourse structure recognition. LEM’s architecture is open for such extensions. With that said, in this paper we have focused on the current stage of development.

LEM will be fully implemented and made available as a web application to the scholarly audience working on Polish. Next, it will be extended with with tools for other languages (e.g. English and German). As LEM has a modular architecture, it would require mostly linking new processing Web Services and adding converters. LEM has an open licences and we will be happy to share our tools, code and *know-how* with teams interested in doing so. Options for exporting to other formats will be added, so that researchers can easily create the output in a particular format (list, text, table) and upload it to other applications (e.g. Mallet) for further processing.

Bibliography

Broda, B., Kędzia, P., Marcińczuk, M., Radziszewski, A., Ramocki, R. and Wardyński, A. (2013). Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. *Studies in Computational Intelligence*. Berlin: Springer, vol. 458, pp. 41-62.

Broda, B. and Piasecki, M. (2013). Parallel, Massive Processing in SuperMatrix – a General Tool for Distributional Semantic Analysis of Corpora. *International Journal of Data Mining, Modelling and Management*, **5**(1):1–19.

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., and Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In: N. Calzolari (ed.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), pp. 1387-1390.

Eder, M., Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: *Digital Humanities 2013: Conference Abstracts*. University of Nebraska-Lincoln, NE, pp. 487-489.

Kędzia, P., Piasecki, M. and Orlińska, M. J. (2015). Word Sense Disambiguation Based on Large Scale Polish CLARIN Heterogeneous Lexical Resources. *Cognitive Studies | Études cognitives*, (15), 269-292.

Kocoń, J. & Marcińczuk, M (2015). Recognition of Polish Temporal Expressions. In Mitkov, R., Angelova, G. & Boncheva, K. (editors), *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 282-290. INCOMA Ltd. Shoumen

Mallet (n.d.) http://mallet.cs.umass.edu/

Marcinczuk, M., Kocon, J. and Janicki, M. (2013). Liner2 - A Customizable Framework for Proper Names Recognition for Polish. *Studies in Computational Intelligence*. Berlin: Springer, vol. 467, pp. 231-253.

Marcińczuk, M. & Radziszewski, A (2013). WCCL Match – A Language for Text Annotation. In Kłopotek, A., M., Koronacki, Jacek, Marciniak, Małgorzata et al (editors), *Language Processing and Intelligent Information Systems*, pages 131-144. Springer Berlin Heidelberg.

PhiloLogi3 (n.d.) https://sites.google.com/site/philologic3/home

Piasecki, M.; Szpakowicz, S.; Maziarz, M. & Rudnicka, E. (2016) plWordNet 3.0 -- Almost There. In Mititelu, V. B.; Forăscu, C.; Fellbaum, C. & Vossen, P. *(Eds.)* Proceedings of the 8th Global Wordnet Conference, Bucharest, 27-30 January 2016, Global Wordnet Association, pp. 290-299.

Piasecki, M., Szpakowicz, S. & Broda, B. (2009). *A Wordnet from the Ground Up*. Wroclaw : Oficyna Wydawnicza Politechniki Wroclawskiej.

Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.

Radziszewski, A. (2013). A tiered CRF tagger for Polish, Intelligent Tools for Building a Scientific Information Platform. *Studies in Computational Intelligence*. Berlin: Springer, vol. 467, pp. 215-230.

Rygl, J. (2014) Automatic Adaptation of Author’s Stylometric Features to Document Types. In Sojka, P., Horák, A., Kopeček, I. and Pala, K. (eds), *Proceedings of 17th International Conference TSD 2014*. Brno, Czech Republic, LNCS 8655, Springer.

Szałkiewicz, Ł. and Przepiórkowski, A. (2012). Anotacja morfoskładniowa. In Przepiórkowski, A., Bańko, M., Górski, R. L. and Lewandowska-Tomaszczyk, B. (eds) (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN., pp. 59-96.

Walkowiak, T. (2015). Web based engine for processing and clustering of Polish texts. *Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX*. Brunów, Poland. Springer, pp. 515-522.

WebSty (n.d.) http://websty.clarin-pl.eu/

Zhao, Y. and Karypis, G. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, **10**(2): 141-168.

Negotiating Meaning and Value: Institutional Research Assistantships, Digital Projects, and Art History

Julia Polyck-O'Neill
jp03uw@brocku.ca
Brock University, Canada

Molly Kalkstein
kalksteinm@ccp.arizona.edu
University of Arizona, United States of America

Introduction

Increasingly, art historians, curators, and other arts scholars in an array of institutions—from galleries, libraries, archives, and museums (GLAMs) to the academy—are turning to digital platforms to facilitate and disseminate new scholarship, promote and provide access to collections, and to engage audiences both within institutional spaces and across geographical networks. Our presentation seeks to use our personal experiences working on digital projects in and related to Digital Art History (DAH) as a mode of analyzing and assessing the current state of the field from the perspective of the paid contract worker. As graduate students with minimal formal training in the Digital Humanities (DH), but with an avid interest in developing our theoretical and methodological knowledge, we feel that our contribution to this emergent and continuously evolving conversation is of value to academic researchers planning and undertaking large-scale digital research projects, particularly those who might be hiring and training research assistants. Because we also seek to contextualize some of the issues we have encountered in the intersections (and interstices) between DAH and DH tools and methods, our study speaks to how these align (or resist aligning) within the conditions of academic institutional realities, where we, as graduate students in Art History, are negotiating the differences between traditional and digital approaches to the discipline.

Project Description

In this paper, we examine how these issues manifest in an on-the-ground institutional context: that of the university research assistantship. We seek to analyze our roles as contract researchers on digital projects, while extending the scope of our examination to include some of the wider-ranging issues that came to light during the course of our work. We then suggest a number of viable solutions and best practices in response to some of the problems we have each encountered, and offer observations about how these issues are systemic within the field of Digital Art History, digital collections, and archives in general—particularly those rich in visual materials.

While traditional Digital Humanities methods work in some instances (for text or data mining, the creation and dissemination of certain kinds of archival or library-based catalogues, GIS-related geographical applications, and so on), we note that DAH—particularly where it engages with a specifically material- or object-based approach—requires its own specialized methods, most of which are still emergent or have yet to be developed. Borrowing from Johanna Drucker's writings on the specific needs of art historians working in digital environments, Diane M. Zorich's studies of extant digital initiatives in cultural institutions in the United States, and bolstered by our own direct experience as graduate students and museum professionals, we perceive that much work needs to be done in the field of DAH.

Perhaps even more importantly, we observe that more effort needs to be directed toward making knowledge of effective DAH methods and platforms with DAH-specific affordances accessible within institutional environments that may not have the infrastructure or expertise in place to carry out DH methodologies, and where digital projects are known to have academic capital yet are not effectually supported. It is our hypothesis that these issues, while generally unintentional, are the result of a maladapted atomization of attitudes and methods that were never intended for the disciplinary particularities of projects dealing specifically with visual, aesthetic, and/or material culture; or for the specific needs of broad public-facing institutions such as GLAMs.

Furthermore despite the gradual emergence of highly visible projects concerning the digital study and dissemination of visual materials—projects such as Object:Photo at the Museum of Modern Art, and the Getty Foundation-funded Online Scholarly Catalogue Initiative—these projects are almost exclusively the purview of large, well-funded institutions. At the same time, we have seen some training opportunities for new digital art historians, such as the 2015 Building a Digital Portfolio institute at George Mason University. As productive as these short-term training institutes are, however, scholars new to DAH require more intensive opportunities to develop these skills, something that many art history departments are currently unprepared to provide. Left out of the equation too are small and mid-sized GLAMs, even those affiliated with large universities, which lack the personnel, experience, and funding to pursue rigorous digital projects. As graduate students and researchers, we have seen first-hand how the absence of a robust network to share ideas and opportunities, teach new skills, and develop new approaches can hamstring the full potential of DAH projects. Although there is a clear interest in and need for platforms and methodologies for Digital Art History, in most instances there is not yet a thriving ecosystem to develop and share them.

Conclusions

Although we offer this as a personal, narratological account of our experiences working as research assistants on digital projects (at Brock University, Ontario; and the University of Arizona, respectively), we also analyze how the programming aspects of these projects may be handled in differing institutional contexts—whether researchers are trained in digital applications, out-of-the-box platforms are adopted, or programming and development are outsourced—and examine these practices in light of their compatibility with both the practical and more abstract aspects of our respective endeavors. We will look at how these decisions inform and shape project development trajectories, as well as how they delimit these projects' longer-term potential. Importantly, we will consider issues of accessibility as they concern research assistants, and examine our own limitations in relation to the usability of the tools and platforms at our disposal in our respective institutions. Additionally, we examine how our independent, unpaid projects allow for the extension of boundaries and exploration in the field, and consider the ethical implications of how unpaid work informs and supports our contributions to funded, institution-based projects.

Bibliography

Drucker, J. "Intro to Digital Humanities." UCLA Center for Digital Humanities. http://dh101.humanities.ucla.edu/?page_id=13

Drucker, J. (2006) "Exhibition Catalogs in the Age of Digital Proliferation." Art on Paper 11, no. 1 (2006): 47-53

Drucker, J. (2006). "Is There a 'Digital' Art History?" Visual Resources: An International Journal Of Documentation 29, no. 1/2 (June 2013): 5-13.

Fisher, M. M., and Swartz, A. (2014). "Why Digital Art History?" Visual Resources: An International Journal Of Documentation 30, no. 2 (2014): 125-137.

Long, M. P. and Schonfeld, R. C. (2014) "Preparing for the Future of Research Services for Art History: Recommendations from the Ithaka S+R Report." Art Documentation: Journal of the Art Libraries Society of North America 33, no. 2 (Fall 2014): 192-205

Zorich, D. M (2012) "Transitioning to a Digital World: Art History, Its Research Centers, and Digital Scholarship." Kress Foundation.org. Last modified May 2012. http://www.kressfoundation.org/uploadedFiles/Sponsored_Research/Research/Zorich_TransitioningDigitalWorld.pdf

Citational Politics: Quantifying Impact in *Digital Scholarship in the Humanities*

Roopika Risam
rrisam@salemstate.edu
Salem State University, United States of America

Amy Earhart
aearhart@tamu.edu
Texas A&M University, United States of America

Introduction

Digital humanities has made an important intervention in scholarly communication, particularly in the realm of citational practices. For example, it has facilitated quantitative analysis of citations within humanities disciplines, illuminated the citational networks in play, and led to the creation of workflows and tools for interpreting citations (Romanello 2016; Crymble and Flanders 2013; Blaney and Meyer 2013; Nyhan and Duke-Williams 2014). Such analysis has much to offer how we understand the confluence of citation, power, and privilege within academic communities of practice.

Yet, the lens of citational analysis has rarely been turned towards digital humanities scholarship itself so we might understand the dominant trends in citational practice. Doing so, however, offers insight into both the citational politics that reinforce homogeneous scholarly practices and illuminates the way that gender, race, and nation are understood in digital humanities. Further, we might map the ways that digital humanities citational practice reveals the contours of the field. One of the few truly international fields, digital humanities moves across nations, languages, and institutional structures. What might citational practices teach us about how digital humanists interact and how areas of inquiry within the field are understood across the world? This line of inquiry answers Isabel Galina's (2013) challenge for greater inclusion in digital humanities scholarship by embracing her mandate:

> We have a combination of scholars who can provide important insights to do this properly. Cultural theory, postcolonial studies, feminist perspectives and other forms of critical theory can make us aware of the problem. But DHers' capacity and willingness to build things can allow us to create projects and tools that help us to be more inclusive.

As such, this paper presents the method, results, and implications of our analysis of citational politics in the journal *Literary and Linguistic Computing* (*LLC*), now *Digital Schol-*

arship in the Humanities (*DSH*). Moreover, we use these results to present recommendations for a new politics of citation that encourages increased diversity of thought and method in digital humanities scholarship.

Analyses of citations in digital humanities scholarship are rare. Domenico Fiormonte has taken up these questions in relation to multilingualism. He argues that Anglophone citations are overrepresented, producing a "monoculture" in digital humanities that devalues scholarly contributions from languages other than English (Fiormonte 2015). Stacy Stutsman (2015) has explored the distribution of digital humanities scholarship in pedagogy, suggesting that the same, narrow list of digital humanities practitioners and theorists from the United States and the United Kingdom - Steven Ramsay, Matthew Kirschenbaum, Lev Manovich, Dan Cohen, Franco Moretti, and Susan Hockey - populate syllabi. These studies raise the important question of which factors shape the citational practices of the digital humanities. While Fiormonte tracks monolingualism in scholarship and Stutsman's study indirectly hints at the effect of nation on pedagogy, the influences of nation, gender, and race on citations in digital humanities journals remain a mystery. These concepts are important, however, because of the reputational and academic currency that citations produce and the growing focus on impact for humanities scholarship. As a result, further analysis of citational practices in digital humanities holds the possibility of uncovering which communities are privileged and disadvantaged by citational practices and how this correlates to other ways of conceptualizing the relationship between access, power, and knowledge in the context of digital humanities scholarship, such as geographical divides between Global North and South and representation within the Alliance of Digital Humanities Organizations (ADHO).

To understand the citational politics of digital humanities scholarship, we began by selecting *DSH* as the subject of our study. Although publications from other constituent organizations of ADHO, such as *Digital Humanities Quarterly* or *Digital Studies / Le Champ Numérique*, are open access and arguably have a significant impact on digital humanities scholarship, we chose to examine *LLC/DSH* because it is the oldest of the ADHO constituent organization's journals, represents the broadest international constituency of digital humanities, and, arguably, is the most influential. We scraped data from the journal to create a data set of titles, authors, affiliations, abstracts, and citations of articles in *LLC/DSH* between 2010 and 2016. We supplemented this data by researching the affiliations of scholars represented in the citation data. We recognize that scholars might move institutions, publish in multiple languages, and reside in different nations. However, we argue that scholarship is formed in particular ways within particular contexts at particular moments in time. Further, this data is not intended to offer precise individual data but larger trends in citation practice.

Using this data set, we examined correlations between the following types of data to identify the primary influences on citational practice:

Language of article
National location of author's institution
Gender

The results of this study show that citation practices converge around subfields within the larger category of "digital humanities" and correlate to both national identity and language. We posit that recognizing such citational practice might help to diffuse tensions between the many methods and approaches that are subsumed under "digital humanities." We also suggest that citational differences are apparent due to gender and language. Based on our results and analysis, we offer concrete solutions for redressing the troubling citational politics evident in our data set.

We will also discuss future steps for this project, including expansion of the data set to include scholarship from *Digital Humanities Quarterly, Digital Studies / Le Champ Numérique*, and the *Journal of the Japanese Association of Digital Humanities.* This will facilitate the exploration of the relationship between paywalled and open access journals and expand the results of our present study. We also intend to undertake a survey to identify racial self-identification of authors and enlarge the data set to explore the influence of race on citational practice and the intersecting influences of race, gender, and nation on the politics of citation in digital humanities scholarship.

Bibliography

Blaney, J., and Meyer, E. T. (2013). "The Problem of Citation in Digital Humanities. *TIDSR: Toolkit for the Impact of Digitised Scholarly Resources.* http://microsites.oii.ox.ac.uk/tidsr/case-study/473/problem-citation-digital-humanities. Accessed October 29, 2016.

Crymble, A., and Flanders, J. (2013). "FairCite." *Digital Humanities Quarterly* 7.2. http://www.digitalhumanities.org/dhq/vol/7/2/000164/000164.html. Accessed October 29, 2016.

Fiormonte, D. (2015). "Towards Monocultural (Digital) Humanities?" *Infolet.* http://infolet.it/2015/07/12/monocultural-humanities/. Accessed October 29, 2016.

Galina, I. (2013). "Is There Anybody Out There: Building a Global Digital Humanities Community." *Red de Humanidades Digitales.* http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/. Accessed October 29, 2016.

Nyhan, J., and Duke-Williams, O. (2014). "Joint and Multiple-Authored Publication Patterns in the Digital Humanities." *Literary and Linguistic Computing* 29.3: 387-399.

Romanello, M. (2016). "Exploring Citation Networks to Study Intertextuality in Classics." *Digital Humanities Quarterly* 10.2. http://www.digitalhumanities.org/dhq/vol/10/2/000255/000255.html. Accessed October 29, 2013.

Stutsman, S. (2015). *Digital Humanities and New Media: Disciplinary Boundaries Explored Through Syllabi.* Poster presented at the meeting of the Humanities, Arts, Science, and Technology Alliance and Collaboratory (HASTAC), East Lansing, MI

Micro DH: Digital Humanities at the Small Scale

Roopika Risam
rrisam@salemstate.edu
Salem State University, United States of America

Susan Edwards
sedwards@salemstate.edu
Salem State University, United States of America

Digital humanities practices are often understood in terms of significant scale: *big* data, *large* data sets, digital humanities *centers* (Terras et al. 2016; Kowalczyk et al. 2014; Borgman 2009; Kretzschmar 2009). This emphasis leads to the perception that projects cannot be completed without substantial access to financial resources, data, and labor (Prescott 2016; Hockey 2016; Evans and Rees 2012). While this can be the case, such presumptions serve as a deterrent to the development of an inclusive digital humanities community with representation across academic hierarchies (student, librarian, faculty), types of institutions (public, private, regional), and geographies (Global North, Global South). In response, how can digital humanities scholars find value in work undertaken at a small scale? This question is at the heart of this paper theorizing the practices of micro digital humanities by reporting on initiatives at Salem State University. These practices include the embrace of minimal computing, small data sets, local archives, and freely available platforms for creating small-scale digital humanities projects while working with undergraduate students.

The work of the Minimal Computing Working Group has articulated a vision for minimal forms of digital humanities praxis (Minimal Computing Working Group 2015). Jentery Sayers (2016) has identified key components of minimal computing, including minimal design, maximum justice, and minimal technical language. These principles privilege access and openness for stakeholders across economic and technical barriers. More importantly, they are precepts that envision how digital humanities practices might be available to those who work outside of macro structures that have historically shaped digital humanities. This has been important at Salem State, a regional, public university undergoing an unprecedented budget crisis due to funding cuts from the state legislature. However, we have faculty and librarians who are committed to using digital humanities to cultivate digital and 21st century literacies in our students. As a result, we have conceptualized a micro digital humanities approach inspired by minimal computing. Micro DH validates scholarly output that does not require digital humanities centers, big data, large data sets, and access to high-performance computing. As an intervention in local digital humanities, it places high value on working with available resources, however small.

At Salem State, we have embraced micro digital humanities through our work with undergraduates. This talk explores these practices in depth, as a model for claiming the legitimacy of small-scale digital humanities. It considers how we have drawn on minimal university resources and existing institutional structures to build a digital humanities community.

First, the focus of our work is our university's archives and special collections, a diverse and free but untapped source of material. This choice emphasizes the primacy of local resources in micro digital humanities. Although Salem is known for its history of the Salem Witch Trials and literature of Nathaniel Hawthorne, our archives focus on the common person's experience in Salem from the mid-19th century to present. This includes a rich history of immigration and activism. It exemplifies the power of micro digital humanities for the diversifying the historical record by giving voice to the ordinary and everyday. Through our work, we shed light on the hidden histories that shape Salem today.

We undertake this work in service of our undergraduate students. Micro digital humanities is an approach that cuts across hierarchies in academic labor, bringing faculty, librarians, and students together to create small projects. Salem State is the most diverse state university in Massachusetts (35% students of color, 40% first-generation college students) and draws a primarily regional, working-class student population. Both the students and the university have few resources, but we work with what we have. To serve this population, we developed the Digital Scholars Program, piloted through a small grant from the university. We designed the program to answer the call of the university's strategic plan to foster student success through research opportunities. Students apply to become Digital Scholars, and those who are selected are mentored through the process of creating a small-scale digital humanities project over the course of a semester.

Because they receive course credit rather than payment, we do not believe that we can, ethically, ask students to work on projects for us. Instead, the program is student-centered and student-led through a scaled down approach emphasizing the creation of micro projects. We select collections for the students to explore related to the history of Salem State then lead them through the experience of creating a digital humanities project from start to stopping point. The process includes making discoveries in the archives, identifying research questions that suit their interests, curating materials, envisioning project design, selecting platforms, and creating a final product - and all the iterative dimensions this process entails. Students also have access to professional development workshops and opportunities to engage with guest speakers who are themselves digital humanities practitioners. Projects our students have undertaken include recovering the history of LGBTQ activism at the university, revealing the colonialist gaze of Salem residents who traveled to India in the 1920s, creating historical models of the university's oldest campus building, and connecting contemporary student activism around Black Lives Matter to the history of anti-racist activism at the university in the 1970s. These projects have helped students engage in a range of practices: digitizing texts, TEI, Omeka, 3D modeling, quantitative textual analysis, data visualization, and oral history. In the spirit of micro digital humanities, we only use freely available resources or open source software we can host ourselves. This is a response to our lack of financial resources but is also a result of our focus on students; we do not want to force them to use proprietary technologies they may not be able to afford to access outside of a university or on their own.

We situate this work as a practice of micro digital humanities, cutting across hierarchies to shift students from the position of consumer of digital media and technologies to the role of producer. This requires setting aside our preferences for what projects based on the collections should look like and recognizing that students will be working at a small scale. However, we view these small projects as pieces of a bigger puzzle that illuminates life in Salem. To bring these projects together, we developed an umbrella digital humanities project called *Digital Salem,* a portal that aggregates student projects by collection. Users visiting *Digital Salem* are offered multiple ports of entry into the history, culture, and literature of Salem. There, the small student projects add up as they contribute to a rich, varied resource on Salem. This experience has suggested how a micro digital humanities can be designed with emphasis connecting small projects as modular pieces that work together to form a bigger picture.

These micro digital humanities practices have been the foundation of the digital humanities community at the university. They have brought together faculty, librarians, and students to facilitate student research at a teaching-intensive university. Further, they offer a model for developing digital humanities at scales appropriate to institutional contexts and strategic planning. Perhaps more importantly, they offer a vision of digital humanities with learning curves and barriers to entry that do not require affiliation with centers, access to expensive technologies, or substantial resources. This, we argue, is essential to the development of an inclusive digital humanities community.

Bibliography

Borgman, C. (2009). The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly* 3.4. http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html. Accessed October 29, 2016.

Evans, L. and Rees, S. (2012). An Interpretation of Digital Humanities. In D. Berry (Ed.), *Understanding Digital Humanities* (pp. 21-41). London: Springer.

Hockey, S. (2016). Digital Humanities in the Age of the Internet: Reaching Out to Other Communities. In W. McCarty & M. Deegan (Eds.), *Collaborative Research in Digital Humanities* (pp. 81-92). London: Routledge.

Kowalczyk, S. T., Sun, Y., Peng, Z., Plale, B., Todd, A., Auvil, L., Willis, C., Zeng, J., Pathirage, M., Liyanage, S., Ruan, G., & Downie, J. S. (2014). Big Data at Scale for Digital Humanities: An Architecture for the HathiTrust Research Center. In W. Hu & N. Kaabouch (Eds.), *Big Data Management, Technologies, and Applications* (pp. 270-294). Hershey, PA: IGI Global. doi:10.4018/978-1-4666-4699-5.ch011

Kretzschmar, W.A. (2009). Large-Scale Humanities Computing Projects: Snakes Eating Tails, or Every End is a New Beginning? *Digital Humanities Quarterly* 3.2. http://digitalhumanities.org/dhq/vol/3/2/000038/000038.html. Accessed October 29, 2016.

Minimal Computing Working Group. (2015). About. *Minimal Computing.* http://go-dh.github.io/mincomp/about/. Accessed October 29, 2016.

Prescott, A. (2016). Beyond the Digital Humanities Center: The Administrative Landscapes of Digital Humanities. In S. Schreibman, R. Siemens & J. Unsworth (Eds.) *A New Companion to Digital Humanities* (pp. 461-475). Malden, MA: Wiley.

Sayers, J. (2016). Minimal Definitions. *Minimal Computing.* http://go-dh.github.io/mincomp/thoughts/2016/10/02/minimal-definitions/. Accessed October 29, 2016.

Terras, M., Baker, J., Hetherington, J., Beavan, D., Welsh, A., O'Neill, H., Finley, W., Duke-Williams, O., Farquhar, A. (2016). Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing, and transforming access to British Library Digital Collections. In *Digital Humanities 2016: Conference Abstracts* (pp. 376-379). Jagiellonian University & Pedagogical University, Kraków.

A Scholarly Edition for Mobile Devices

Peter Robinson
peter.robinson@usask.ca
University of Saskatchewan, Canada

Barbara Bordalejo
barbara.bordalejo@kuleuven.be
University of Leuven, Belgium

It has frequently been asserted that the digital turn may make it possible to bring scholarly materials, and specifically literary texts, to new and larger audiences than ever before possible (e.g. Jewell 2009). The new breed of mobile device (tablets and smartphones), combining ease of use and powerful interfaces, present an extraordinary opportunity, to make new kinds of books to reach new readers. However, almost all digital books so far made conform to the Erin McKean's characterization of them as "paper thrown on a computer screen": thus eReaders and .pdf files (2007). It is clear that as producers of eBooks, we are achieving less than we could. In this paper, we survey attempts to produce digital works which both reach new audiences and which offer new perspectives on reading.

A few publishers have produced exemplary digital representations of literary texts which have shown, decisively, that digital books can be far more than eReaders and pdfs. The TouchPress publications of T. S. Eliot's *The Waste Land* and Shakespeare's Sonnets, among others, show how reading in the digital medium can be embedded in rich structures of image and sound, extending far beyond the printed page. TouchPress has been extraordinarily successful in reaching huge audiences, reporting in 2012 over 290,000 purchases of its "The Elements" app, and over half a million sales of all its apps (*The Guardian* 2012). The Barcelona-based "Play Creativad" team has shown too with their iClassics publications how much can be achieved with imagination and rather minimal resources. The original release of iPoe has been downloaded more than 500,000 times. These publishers have proved that classic texts may be given new life within the App environment.

These publications have, however, been the exception rather than the rule. Between them, TouchPress and the Play Creativad team have published fewer than ten literary texts. Further, the resources (in the case of TouchPress) and creative ingenuity (in the case of Play Creativad) underpinning these books are in short supply. In this paper, we report on our own attempts to follow the path blazed by these two in our own work on Geoffrey Chaucer and the *Canterbury Tales*. We are acutely aware that we lack both the resources and creative force of our models. On the other hand, we know our text very well, and have scholarly resources to deploy. This paper will describe the CantApp: a new way of presenting Chaucer's *Canterbury Tales*, beginning with the "General Prologue." The CantApp breaks new ground in several respects. It is the first App (an application specifically designed for tablet computers and other mobile devices) to present an authoritative edition of the *Canterbury Tales*, or indeed, to our knowledge, any medieval English text. Second, it will target an audience usually not specifically addressed by scholarly publications: students and other readers encountering Chaucer for the first time. Third, it will present a new reader's text of the *Tales*, based on the work of the Canterbury Tales Project and designed to allow new readers to discover for themselves the *Tales* in the most original form now available. Fourth, while designed primarily for beginning readers of Chaucer, it will be created to the highest scholarly standards, and will incorporate materials usually found only in advanced research-oriented publications, notably Richard North's annotations and explanations. Fifth, and most important: our aim is to use this chance not just to allow students to read Chaucer as they might read him in existing print editions, but to read him in a new way.

We aim not just to bridge the gap between digital editions containing little, but read by many, and those containing much, but read by a few. Our aim is to enable a new way of reading Chaucer. We contend (as have many before us) that the *Canterbury Tales* is a multimedia object containing both visual and aural elements. Visual: in that the poem comes to us from a manuscript culture in which the marks on the page – the ornamentation, the use of emphatic devices, the paratext, the forms of writing – are indicators of meaning. Aural: in that the poem comes alive from the page when it is read aloud and heard, in performance. Accordingly, the central axis of the CantApp is the simultaneous presentation of the key Hengwrt manuscript of the *Tales*, in high-resolution full-colour digital images, synchronized to the sound text, itself very close to the Hengwrt manuscript. Thus, the reader will hear the words, spoken in our best effort at the original pronunciation, with poetic and dramatic expression, with each corresponding line of the manuscript scrolled and highlighted insync, and with glosses, explanations and notes always one click away. All we want to offer is already available on the internet of course: but always at a distance, so that the reader is distracted away from the text. By centring the reader's experience on the sound of the text and the appearance of the manuscript, supplemented as needed, we can bring Chaucer to beginning readers with a rich immediacy hitherto available only after years of study.

As well as presenting our own new reader's text, the manuscript, and performance in sound, the CantApp includes Terry Jones' "minimal translation" of the General Prologue, with his annotations. Jones is uniquely qualified as a reader of the *Tales*: not only by his own formal academic writings, but through the transmutation of Chaucer which has fed his own creative work for nearly half a century. How Terry Jones reads Chaucer, word by word and line by line, is revelatory. Hence, the reader of the CantApp can compare his or her own developing sense of Chaucer with that of Terry Jones. In addition, the creators of the CantApp will provide all the materials the beginning reader will need: annotations, textual notes (with further manuscript images), glosses of difficult words.

It is a conventional wisdom, that – several decades into the digital revolution --traditional print books still remain the best way to read the *Canterbury Tales*, and indeed for any major work of literature. The new generation of digital books, as those mentioned from TouchPress and Play Creatividad, is challenging that wisdom. A key factor is the App environment. Unlike a PDF reader, or the Kindle or similar paper-on-screen systems, a well-made App is not a surrogate for a print book or anything else: it is a new kind of experience. Unlike reading in internet browsers, where any number of factors (unpredictable screen sizes, network failures) may disrupt the reading experience, the App environment permits complete control over exactly what the reader sees, and instantaneous response to reader actions. The key word is immediacy: within an App, the reader may move as surely and rapidly from one part of the publication to any other as easily as one may turn the page in a printed book.

Furthermore, the multimedia possibilities of an App provide opportunities for reader engagement no book can match. The performative aspect of the *Tales* may be directly expressed through the reading we offer, and the line-by-line synchronization of the reading with the manuscript, the edited text and translation, offer an immersive reading experience fed by image and sound as well as words on the page. Our hope is that readers will find themselves as completely engaged with the *Tales*, though in a different key, as could be achieved by any printed book. Indeed, the judicious matching of text and sound with notes and glosses (prepared by Richard North) designed to inform without distracting might offer beginning readers a fast and accessible way into the *Tales*.

The resources needed for Apps constructed according to this template -- images of the original source, a sound edited text, a reading of the text, translation and annotations -- may reasonably be gathered by many scholarly projects. For making the App, we used the well-known, robust and (importantly!) free PhoneGap system, and making the whole required only a few weeks of programmer time (including learning how the PhoneGap worked). Because of these low marginal costs, we are able to make the App available at low cost (not more than 1 euro) or free. We are considering following the Play Creativad model, of selling at low cost most of the time, but with significant periods of free distribution. The measure of our success will be not just how many readers we get for this publication, but the number of other projects which follow this model.

Bibliography

Dredge, S. (2012) "Touch Press passes 500k book-app sales milestone on iPhone and iPad", *The Guardian* 20 July 2012, https://www.theguardian.com/technology/appsblog/2012/jul/20/touch-press-book-apps-success Accessed 1st November 2016.

Jewell, A. (2009) "New Engagements with Documentary Editions: Audiences, Formats, Contexts." *Library Conference Presentations and Speeches*. The Libraries at University of Nebraska-Lincoln, http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1056&context=library_talks. Accessed 1st November 2016.

McKean, E. (2007) "The Joy of Lexicography." TED. March 2007. Lecture.

Poe, E. A. (2013). iPoe. Barcelona, PlayCreatividad

Playful Instruments: Reimagining Games as Tools for Research and Scholarly Communication

Jon Saklofske
jon.saklofske@acadiau.ca
Acadia University, Canada

In academic contexts, digital games are often studied as texts or are used as pedagogical tools to teach basic concepts in early education situations. Less usefully, their systems and economies are often co-opted and decontextualized in short-sighted attempts to "gamify" various aspects of learning or training. It's no wonder that digital games often appear as a marginal endeavor in Digital Humanities practices, despite their relative compatibility with the broad scope of DH methods and perspectives. However, given that games are highly controlled, conditional, choice-and-consequence-based, problem-solving environments in which players are expected to interact with simulated settings and elements after agreeing to take on particular roles and subject positions, there are promising potential uses of these experiences in academic contexts that have not been fully considered.

One of the unique ways that DH provocations and practices have managed to create a critical lucidity is through making, through the construction of prototypes and through discussions regarding alternative models of perception, narrativity, organization and understanding, enabled through the computer's multi-media frame. DH work often defamiliarizes and repoliticizes the forms and functions of communication and scholarly work, and continues to challenge socially constructed and sustained institutional habits by asking "What is the relationship between making, thinking, using and critique?" "Thinking through" tools, prototypes, interfaces and platforms and through the narratives that such processes construct, DH-inspired experiments are slowly, but noticeably and provocatively expanding opportunities for scholarly research and communication methods and means. Such changes are the product of an imaginative resistance to traditional limitations and habitual practices in our institutions.

Motivated by the imperative to explore alternative modes and methods of scholarly research and communication, and guided by the values of open social scholarship practices, this paper reconsiders games not as things to study, but as instruments to study with. Given that games can function as simulations, models, arguments and creative collaboratories, game-based inquiry can be used as a potential method of humanities research and communication. While these ideas have been explored in a preliminary way in relation to a few different academic disciplines (Donchin, 1995; Boot, 2015; Mitgutsch and Weise, 2011; Westecott, 2011) this paper will make the case for a humanities-calibrated consideration of the pragmatics and potentials of game-based research, games as instances of critical making, critical intervention and scholarly communication, and more complex forms of game-based learning than those currently practiced.

This paper is not about shifting the focus of existing game paradigms and practices to more productive and instructional/educational ends (as is often done in "serious game" or "edutainment" design). It is an attempt to challenge textually-dependent scholarship with game-based processes while simultaneously challenging conventional game features and functions with scholarly creativity and textual affordances. In other words, I am concerned with renewing the process of scholarly inquiry in the humanities via game creation and experience, akin to the richness of what Joanna Drucker describes as "diagrammatic process." I have also been inspired by Geoffrey Rockwell's attempt to avoid the term "serious games" while also promoting the idea that developing and playing through games are viable ways of modeling and reflecting on humanities based research activities.

Several examples will be discussed, including the use of the open-source Twine program in my undergraduate and graduate classes as a simple and accessible game-design engine. Twine can be used to construct environments which ask critical, provocative questions, or which attempt to rhetorically persuade players through interactive experience and interpellative role-playing. Anna Anthropy's *Queers in Love at the End of the World*, Porpentine's *Those We Love Alive*, Mattie Brice's *Blink*, Kitty Horrorshow's *Daymare #1: Ritual,* Pippin Barr's *Burnt Matches*, and Zoe Quinn's *Depression Quest*, are examples of Twine being used as an innovative and disruptive game engine. In addition, Merrit Kopas' book, *Videogames for Humans*, thoroughly explores and justifies the diversity of game-based experiments in Twine: "Authors are doing things with Twine that aren't possible with traditional text. And at the same time, they're using interactive media to tell stories that mainstream videogames couldn't dream of telling" (Kopas, 2015: 11). My students have used Twine to engage their peers in alternatives to essay communication, and working together to design such experiences involves them in a collaborative form of critical making. Composing a networked narrative in Twine is akin to constellating and curating not only ideas, but multiple pathways through such ideas. This critical mapping process is as important as the selective routing process experienced by players, who trace particular storylines through the environment.

The usefulness of such methods can be demonstrated through an assignment submitted by Rebecca Wilson, one of my graduate students, who used the creative process of designing a Twine-based experience to achieve the following research goals:

1. To better comprehend the relationship between William Blake's creative process, his biographical context and his prophetic works,
2. To model and critique his complexity through an emulation of Blake's own disregard for temporal and spatial consistency and his transitional unpredictability, and
3. To explore the Twine engine as a site of utopian hope, utopian method and heterotopic tensions (thus responding to and engaging with theoretical ideas advanced by Michel Foucault (1984) and Ruth Levitas 2013)).

Addressing these questions through the methodology of building the gamespace and producing a written reflection on the process is different from the resulting game experience in which players role-play as William Blake, interactively negotiating and determining causal links between everyday experiences, inspirational visions and creative invention. However, both opportunities demonstrate the variety of ways in which games, game engines, and game platforms can be used as instruments for research, scholarly communication and pedagogy. As well, given that Twine's output is an HTML file that can be served online and accessed through a web browser on multiple devices, Wilson's work is now accessible to a much broader audience than an academic paper on the same topic.

Another example of the ways that games can be used as research tools is an in-progress project that is looking to generate feminist game prototypes to facilitate new models, diverse approaches and different narratives to expose war's patriarchal morphology, to provoke a rethinking of militarised masculinity in the real world and the persistence of hyper-masculinity in idealistic representations of war. These games, designed with alternative value sets to distinguish them from traditionally masculine power fantasies are meant to challenge and realign the values that players become reflexively used to perceiving and employing within mediated scenarios of militarisation and armed conflict. The adoption of feminist value-based design challenges habitual idealisations of war, violence and hyper-masculinity in video game environments through feminist perspectives. Multiple and contradictory feminist war game prototypes can be used to disrupt habits of institutional/personal perception and practice and provocatively/performatively engage players with complex issues of violence, gender and media culture. Incorporating feminist values into a game's design creates an intervention that promotes critical lucidity for players.

Identifying the need to confront and challenge traditional habits of head, hand, heart and media representation when it comes to game-based perceptions of war (which can reflect and configure attitudes towards war in general) is not unique to this talk. Mary Flanagan, in "Practicing a New Wargame" discusses the perceptual limitations reproduced by conventional wargames and calls for alternative ways of imagining conflict resolution:

> "We must look to transcend old conflict models, or we risk perpetuating the damaging myth that there are limited ways of resolving conflicts…. It is vital that game scholars, makers and players see these familiar models on a continuum of change, so new play forms that model new solutions to our problems can be invented. Our games are constantly evolving, and this means we all have an opportunity, even a responsibility, to evolve with them and push ourselves to model the world we wish to create." (2016: 706)

Flanegan's call is an important acknowledgement of game design and gamespace as an opportunity to imagine new models. An effort to reimagine, pluralize and critically engage with the ways that we realize and idealize the long history of armed conflict via a critique of naturalized perceptual habits and assumptions directly speaks to the broader need to expose, augment and erode habitual ways of seeing, being, doing and narrating in the humanities.

These initial examples justify the use of games as research and scholarly communication methods, as more than just new media texts to interpret, and aligns them with current DH efforts to use computer technology as interventions to challenge, critique and re-humanize systems, ideologies, and habitual narrativities, to pluralize perspectives, confront complexity and facilitate multiple models of perception and practice. The computer is a flexible tool can be used in diverse ways to broaden our understanding of human culture and to generate inclusive, inhabitable and thought-provoking stories. By foregrounding the values of open social scholarship and engaging with broader publics via mechanical extensions of perception and action, these unconventional approaches work in connective, integrative and expansive ways to avoid modelling humanities scholarship on more conventional game mechanisms and goals that unproductively foreground acquisition, exploration and competition.

Bibliography

Boot, W. R. (2015). "Video games as tools to achieve insight into cognitive processes." *Frontiers in Psychology*, 6: 1-3.

Donchin, E. (1995). "Video games as research tools: The Space Fortress game." *Behavior Research Methods, Instruments, Computers*, 27(2): 127-53.

Drucker, J. (2013). "Diagrammatic Writing." *new formations: a journal of culture/theory/politics*, 78: 83-101.

Flanagan, M. (2016). "Practicing a New Wargame." In *Zones of Control: Perspectives on Wargaming.* Eds. Pat Harrigan, Matthew G. Kirschenbaum & James F. Dunnigan. Cambridge: MIT, pp. 703-708.

Flanagan, M., and Nissebaum, H. (2014). *Values at Play in Digital Games.* Cambridge: MIT.

Foucault, M. (1984). "Of Other Spaces: Utopias and Heterotopias." Trans. Jay Miskowiec. *Architecture/Mouvement/Continuité*, pp. 1-9.

Kopas, M., ed. (2015). *Videogames for Humans*. Instar.

Levitas, R. (2013). *Utopia as Method*. Basingstoke: Palgrave MacMillan.

Mitgutsch, K. and Weise, M. (2011). "Subversive Game Design for Recursive Learning." *DIGRA 2011 Proceedings: Think Design Play*.

Rockwell, G. (2003). "Serious Play at Hand: Is Gaming Serious Research in the Humanities?" *Text Technology,* 2: 89-99.

Westecott, E. (2011). "Games as Research Tools" *Vimeo*, uploaded by Toronto Digifest, 15 Dec., https://vimeo.com/33724936.

Stable Random Projection: Standardized universal dimensionality reduction for library-scale data

Benjamin Schmidt
bmschmidt@gmail.com
Northeastern University, United States of America

Summary

This paper describes a new method for dimensionality reduction, "stable random projection," (hereafter "SRP") distinctly suited for large textual corpora like those used in the digital humanities. The method is computationally efficient and easily parallelizable; scales to the largest digital libraries; and creates a standard dimensionality reduction space for all texts so that corpora and models can be easily exchanged. The resulting space makes a wide variety of applications suitable to bag-of-words data, such as nearest neighbor searches, classification, and semantic querying possible with data sets an order of magnitude smaller in size than traditional feature counts.

SRP is a minimal, universal dimensionality reduction with two distinctive features:

1. It makes *no distinction between in- and out-of-domain vocabularies.* In particular, unlike standard dimensionality reduction it creates a single space that can hold documents of *any language*.
2. It is *trivially parallelizable*, both on a local machine and through web-based architectures because it relies only on code that can be easily transferred across servers, rather than requiring large matrices or model parameters.

These two features allow dimensionality reduction to be conceived of as a piece of infrastructure for digital humanities work, rather than just an ad-hoc convention used in a particular project. This method is particularly useful for provisioners and users of text data on extremely large and/or multilingual corpora. This creates a number of new

applications for dimensionality reduction, both in scale and in type. SRP features could usefully be distributed by libraries as a (much smaller and easier to work with) supplement to feature counts. After a description of the method, some novel uses for dimensionality reduction on such libraries are shown using a sharable dataset of approximately 4,500,000 books projected into SRP-space from the Hathi Trust.

Description of the method

The goal of SRP is to reduce of text of uncertain length to a much smaller fixed- length vector to which the many tools of textual analysis, machine learning, and linear algebra can be applied. The core technique used here for dimensionality reduction is *random projection.* Random matrix theory has emerged in the past few decades as an useful alternative to more computationally complex forms of dimensionality reduction.(Halko, Martinsson, and Tropp 2009) I make use here of the observation that it is possible to project into a space where points as determined purely by sampling randomly from the set [-1,1].(Achlioptas 2003) A true random number generator is not suitable for reproduction. The other core element of SRP, therefore, is a quasi-random projection for every individual word created using cryptographic hashes (specifically, SHA-1).

This allows the method to be defined algorithmically, making it easy to apply to any text. I have written short code libraries to implement the transformation in the three most important language for DH tool development: Python, R, and Javascript. These include a few necessary additional conventions such as minimal tokenization rules, a method for expanding beyond the 160 dimensions provided by SHA, and the byte-encoding of the Unicode character sets.

Comparison to existing methods

The gold standard for dimensionality reduction are techniques that make use of co-occurrences in the term-document matrix such as latent semantic index- ing and independent components analysis. More recent techniques such as semantic hashing can be even faster and more efficient at optimally organizing documents in various types of vector spaces designed especially for particular documents.(Salakhutdinov and Hinton 2009) Another strategy finding recent use in the digital humanities is using an LDA topic model as dimensionality reduction, which produces neatly interpretable dimensions for analysis (Schöch, 2016; Fitzgerald,2016). In both the digital humanities and computer science, scholars frequently use "top-N" words as a good enough approximation of the textual footprint, limiting the dimensions to a few hundred of the most common words in the corpus, producing what Maciej Eder has called "endless discussions of how many frequent words or n-grams should be taken into account" for stylometry.(Underwood 2014, Eder (2015))

These methods suffer two problems that make them problematic as a *general-use* feature reduction. First, the better ones are computationally complex, and quite difficult to perform on a very large corpus. Second, it is difficult or impossible to project *out-of-domain* documents into the space from a standard projection if they contain vocabulary different than the training corpus. This out-of-domain problem presents a particularly great problem for multilingual corpora, because texts that are missing or in sparsely-represented languages will behave erratically in the new environment.

Some other work in the digital humanities and computer science has used hashes, random projection, and other similar methods as an ad-hoc rather than infrastructural technique. SRP can be thought of as a particular species of *locality-sensitive hashing*, another version of which has been used by Douglas Duhaime to identify reuse in poetic texts based on three-letter phrases.(Duhaime 2016). Also related is the "hashing trick" in computer science(Weinberger et al. 2009), which is better than SRP in many ways for the short documents computer scientists frequently study, but takes significantly more memory to store for book-length documents (an edge case in the computer science literature, but among the most important for humanists).

Applications

This reduced space can be put to many of the same uses as a standard bag-of- words model in considerably less space and with the potential for building web facing tools. Among those to be described are:

1. **Duplicate detection**. SRP is quite accurate at identifying duplicate books in a computationally tractable space using cosine similarity, both inside a corpus and across disparate corpora.
2. **Similarity Search**. A prototype web page allows any user to paste in any text; it will hashed on the client side into the standard space, and a server can return in a few seconds the most similar documents. The top entries can function for duplicate detection; the lower ones presenting interesting opportunities for exploratory analysis. A search for *Huckleberry Finn*, for example, finds a large number of other American adventure novels about boys in the American west.
3. **Classification**
   - SRP features perform approximately as well as top-n words (~77%) on a pre-existing task described by Ted Underwood, separating high- from low-prestige poetry.(Underwood 2015)
   - A single hidden layer neural network trained with 640-dimensional SRP features can accurately classify a held-out sample of books into one of 225 Library of Congress Classification subclasses (for example, whether a work is PR: British Literature or PS: American Literature) with ~78% accuracy based on about 1 million training examples. A single classifier works in multiple languages simultaneously; its determinations on arbitrary pasted

text are accessible for inspection through a web site.

- A different single hidden layer neural network trained with SRP features and a novel encoding scheme for years using Google's TensorFlow framework can accurately predict the years for withheld books with a median error of four years from the true publication date.

SRP as Access

SRP fits in the DH2017's theme of "Access" in two ways.

First, it makes many forms of text analysis on huge digital libraries far more feasible for scholars without access to high performance computing resources. On large corpora, data storage and dimensionality reduction can be more resource- intensive than the actual analysis. The dimensionality-reduced dataset for the full Hathi Trust corpus can fit into 10 GB, easily storable on most computers; subsets are suitable for use in classroom or workshop settings.

Second, the ease with which it works with distributed web architectures, and its language agnosticism, can create new routes into neglected portions of large archives, particularly those with insufficient metadata.

Bibliography

Achlioptas, D. (2003). "Database-Friendly Random Projections: Johnson- Lindenstrauss with Binary Coins." *Journal of Computer and System Sciences,* Special issue on PODS 2001, 66 (4): 671–87. doi:10.1016/S0022-0000(03)00025-4.

Duhaime, D. (2016). "Plagiary Poets. Plagiary Poets." http://plagiarypoets. io/.

Eder, M. (2015). "Visualization in Stylometry: Cluster Analysis Using Networks." *Digital Scholarship in the Humanities,* November, fqv061. doi:10.1093/llc/fqv061.

Fitzgerald, J. D. (2016) "What Made the Front Page in the 19th Century?: Computationally Classifying Genre in 'Viral Texts". July 13 2016 http://jonathandfitzgerald.com/blog/2016/07/13/keystone-paper.html

Halko, N., Martinsson,P.-G., and Tropp, J. A. (2009). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Ap- proximate Matrix Decompositions." arXiv:0909.4061 [Math], August. http://arxiv.org/abs/0909.4061.

Salakhutdinov, R., and Hinton, G. (2009). "Semantic Hashing*." International Journal of Approximate Reasoning,* Special section on graphical models and information retrieval, 50 (7): 969–78. doi:10.1016/j.ijar.2008.11.006.

Schöch, C. (2016, pre-publication) "Topic Modeling Genre: An Exploration of French Classical and En- lightenment Drama"*. Digital Humanities Quarterly.*

Underwood, T. (2014). "Understanding Genre in a Collection of a Million Volumes, Interim Report." http://figshare.com/articles/Understanding_Genre_ in_a_*Collection_of_a_Million_Volumes_Interim_Report*/1281251.

Underwood, T.. (2015). "The Literary Uses of High-Dimensional Space." *Big Data & Society* 2 (2): 2053951715602494. doi:10.1177/2053951715602494.

Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). "Feature Hashing for Large Scale Multitask Learning." In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1113–20. ICML '09. New York, NY, USA: ACM. doi:10.1145/1553374.1553516.

Phygital Augmentations for Enhancing History Teaching and Learning at School

Susan Schreibman
susan.schreibman@gmail.com
Maynooth University, Ireland

Constantinos Papadopoulos
cpapadopoulos84@gmail.com
Maynooth University, Ireland

Brian Hughes
brian.hughes@nuim.ie
Maynooth University, Ireland

Neale Rooney
neale.rooney@gmail.com
Maynooth University, Ireland

Colin Brennan
colin.mark.brennan@gmail.com
Maynooth University, Ireland

Fionntan Mac Caba
fionntanmaccaba@gmail.com
Maynooth University, Ireland

Hannah Healy
healyha@tcd.ie
Maynooth University, Ireland

Contested Memories: The Battle of Mount Street Bridge (BMSB), funded by the Andrew W. Mellon Foundation as part of the Humanities Virtual World Consortium, is an online and annotated digital reconstruction in Unity 3D of a seminal battle of the Irish Easter Rising of 1916. A goal of the project was to leverage phenomenologies of time and space as afforded by virtual world technologies to provide new insights and help to answer what had heretofore been intractable questions about how the battle unfolded. The battle, as well as the 3D reconstruction, received a great deal of attention during the Centenary of the Easter Rising, including the specially commissioned radio documentary (Documentary on One, RTE Radio 1) *Battle at the Bridge*.

Building on the interest in and success of this project, the project team was encouraged to further develop the project so that it could be used in secondary schools. Due to the exigencies of technology in Irish classrooms, including limitations on booking lab time with many computers too old to run the VR software, it was decided that a different approach would be taken. Hence the BMSB Augmented Reality mobile app was developed (figure 1), to educate students about the facts of the battle, while providing a deeper and more holistic understanding of war and its effects. At the same time, it provided the project team with the opportunity to leverage digital humanities research within a participatory engagement setting to reach out to second level students and teachers to provide them with opportunities of engaging with cutting edge technologies integrated into a student-led learning environment.

This project shares a philosophic approach with other maker projects in the Digital Humanities, as Sayers et al. (2016), quoting Neil Gershenfield, describe in *A New Companion to Digital Humanities* as "the programmability of the digital worlds we've invented" applied "to the physical world we inhabit" in which objects "move easily, back and forth, in the space between bits and atoms." Thus Augmented Reality (AR), a technology that superimposes digital content, including images, animations, and annotations, over the real world, was decided on as the key technology to translate our research into a classroom setting. Previous research on the affordances of AR for teaching and learning have highlighted that interactivity, collaboration, problem-solving, and narratives mediated through technology can aid both engagement and understanding (Dunleavy et al., 2009). Several AR applications have been developed for primary and secondary education for fact-based topics, such as geometry, astronomy, chemistry, and the human body (see for example Chromville Science; Anatomy 4D; Elements 4D. However, AR applications for humanities subjects are limited. To the authors' knowledge there is no history-based AR application that is designed for use in the classroom, although there are several mobile-based student-centred AR for outdoor historic sites (see for example: Schrier, 2005 for the Battle of Lexington and Singh et al., 2014 for the Christiansburg Institute).

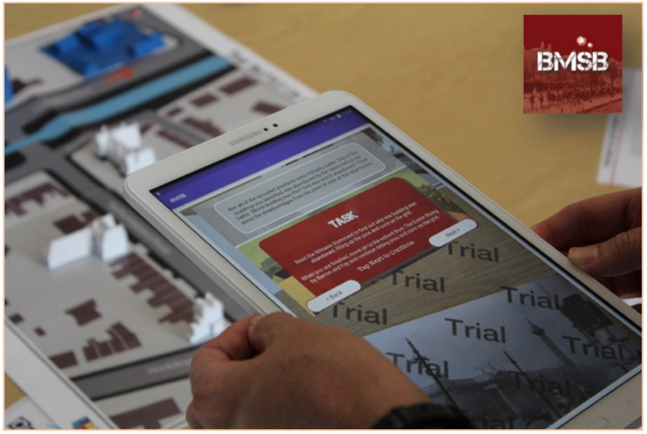


Figure 1. Early prototype of the BMSB Augmented Reality app

These kinds of technologically-driven blended-learning applications which improve digital literacy are seen as the key to the future of education, not only in Ireland, but abroad (see the Digital Strategy for Irish Schools, 2015-2020). Their employment might help to solve challenges such as personalised learning, keeping education relevant, and blending formal and informal processes (Johnson et al., 2016). This project also resonates with the call for a change in history instruction; from one that promotes passive consumption of facts to that which makes more use of analogue and digital primary sources to foster historical thinking (Tally and Goldenberg, 2005; Lee et al., 2006; Stripling, 2011), teaches historical reasoning (van Drie and van Boxtel, 2007), and provides the opportunity to problematise different sources, evaluate arguments, and form new interpretations (Borton and Levstik, 2004).

In terms of app design, the project team were aware of the challenge of how technologies have shifted our attention from the physical to the digital space: interactions with devices absorb attention, distracting from physical experiences and social interaction (Chrysanthi et al., 2012). Therefore, a premise of the project is that, if digital technologies are to assume an active role in history teaching and learning, ways to actively engage students by creating conditions that blend the physical and the digital through participatory and hands-on engagements need to be employed (figure 2).

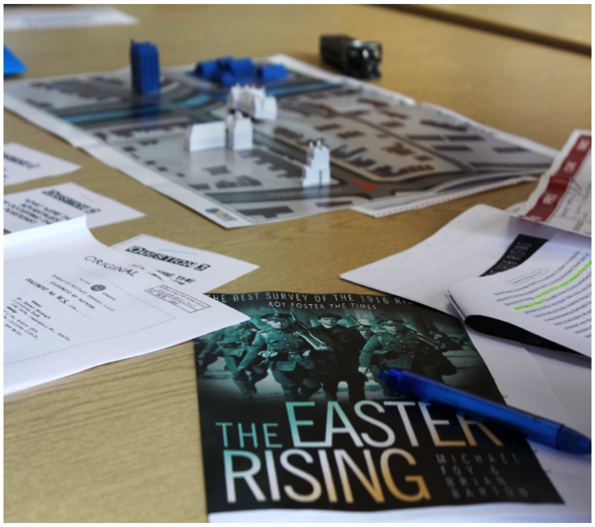


Figure 2. Physical Materials used in 'Lesson 1 - Group 1: The Buildings

The BMSB AR app (developed for Android tablets using the Wikitude Software Development Kit for Android Studio) employs a phygital approach by enabling a task-based digital and AR exploration triggered by physical objects and primary sources, including photographs, witness statements, 3D printed buildings, and state records (figures 2, 3). Since these have a central role in the lessons, students have to evaluate their content, origins, authority, and reliability and finally through group work to form their own interpretations.

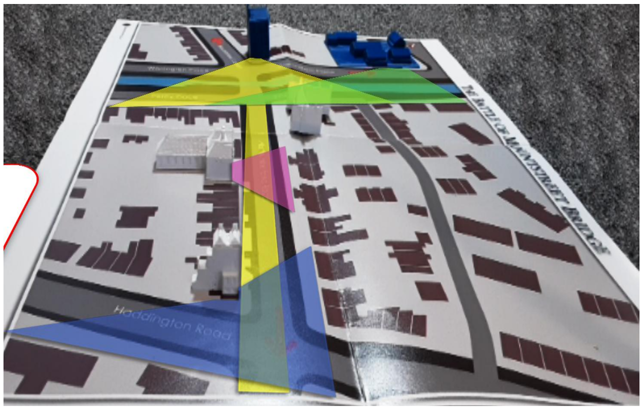


Figure 3. AR element of the app showing over the physical map the lines of fire from the four buildings that were occupied by the Irish volunteers

This project followed an iterative design methodology. The first two focus groups were carried out with second level history teachers (figure 4) to ascertain a) logistical information about schools (e.g. wireless, class size, classroom arrangement, labs etc.), and b) how to develop content and design interactions in order students to get the most out of this experience. The results of the focus group made it clear that due to the wide variance in technology in schools across Ireland, the project had to follow the paradigm of handling/ activity boxes in museums (Comer, 2014) in which everything the teacher needs would be posted, including the tablets, physical materials, and lesson plans for a period of two weeks.

Figure 4 Second Level history teachers testing the AR app during a focus group in August 2016

This paper will report on the different stages of development of the AR app, the difficulties encountered as well as strengths of the approach. It will also describe the results of the testing and formal evaluations carried out with secondary school students and the next stages for the release of the prototype across Ireland.

Bibliography

Anatomy 4D (2016). Daqri. http://anatomy4d.daqri.com/ (Accessed 30/10/16)

Barton, K., and Levstik, L. (2004). *Teaching History for the Common Good*. Mahwah, New Jersey: Lawrence Erlbaum.

Battle at the Bridge (2016). Documentary on One. RTE Radio 1. http://www.rte.ie/radio1/doconone/2016/0308/773342-battle-at-the-bridge/ (Accessed 30/10/16)

Chromville Science (2015). https://chromville.com/chromvillescience/ (Accessed 30/10/16)

Chrysanthi, A., Papadopoulos, C., Frankland, T., and Earl, G. (2013). "'Tangible Pasts': User-centred Design of a Mixed Reality Application for Cultural Heritage" In Earl, G. et al. (eds) *Archaeology in the Digital Era. Papers from the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA)*, Southampton, 26-29 March 2012. Amsterdam: Amsterdam University Press, pp. 31-41.

Comer, L. (2014). *Bronze Age Handling Box Resource Book*. National Museum of Ireland.

Digital Strategy for Schools 2015-2020 (2014). *Enhancing Teaching, Learning and Assessment. Department of Education and Skills*. Online: https://www.education.ie/en/Publications/Policy-Reports/Digital-Strategy-for-Schools-2015-2020.pdf (Accessed 28/10/16)

Dunleavy, M., Dede, C. and Mitchell, R. (2009). "Affordances and Limitations of Immersive Participatory Augmented Reality Simulations for Teaching and Learning." *Journal of Science Education and Technology*, 18: 7. doi:10.1007/s10956-008-9119-1.

Elements 4D (2016). Daqri. http://elements4d.daqri.com/ (Accessed 30/10/16)

Gartner (2016). *Hype Cycle for Emerging Technologies*. Online: http://www.gartner.com/newsroom/id/3412017 (Accessed 29/10/16)

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., and Hall, C. (2016). *NMC Horizon Report: 2016 Higher Education Edition*. Austin, Texas: The New Media Consortium. Online: http://cdn.nmc.org/media/2016-nmc-horizon-report-he-EN.pdf (Accessed 28/10/16)

Lee, J. K., Doolittle, P. E., and Hicks, D. (2006). "Social studies and history teachers' uses of non-digital and digital historical resources." *Social Studies Research and Practice*, 1(3): 291-311.

Sayers, J., Devon, E., Kraus, K., Nowviskie, B., and Turkel, W. J. (2016). "Between Bits and Atoms: Physical Computing and Desktop Fabrication in the Humanities." In Schreibman, S., Siemens, S., and Unsworth, J. (eds) *A New Companion to Digital Humanities.* London: Wiley-Blackwell, pp. 3-21

Schrier, K. (2005). *Revolutionizing History Education: Using Augmented Reality Games to Teach Histories*. PhD Thesis. Massachusetts Institute of Technology. Online: http://cmsw.mit.edu/revolutionizing-history-education-using-augmented-reality-games-to-teach-histories/ (Accessed 28/10/16)

Singh, G., Bowman, D. A., Hicks, D., Cline, D., Todd Ogle, J., Johnson, A., Zlokas, R., Tucker, T. (2015). "CI-Spy: Designing A Mobile Augmented Reality System for Scaffolding Historical Inquiry Learning." In *IEEE International Symposium on Mixed and Augmented Reality - Media, Art, Social Science, Humanities and Design*, pp.9-14. doi: 10.1109/ISMAR-MASHD.2015.19 http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7350728 (Accessed 28/10/16)

Stripling, B.K. (2011). *Teaching the Voices of History through Primary Sources and Historical Fiction: A Case Study of Teacher and Librarian Roles.* Syracuse University. PhD Thesis.

Tally, B., and Goldenberg, L. B. (2005). "Fostering historical thinking with digitized primary sources." *Journal of Research on Technology in Education*, 38(1): 1-21.

van Drie, J. and van Boxtel, C. (2007). "Historical reasoning: Towards a framework for analyzing students' reasoning about the past." *Educational Psychology Review*, 20: 87-110.

The Role of Digital Humanities in Countering Cultural Genocide: the virtual reconstruction of Julfa Cemetery

Harold Short
haroldshort@mac.com
Australian Catholic University, Australia

The trail of destruction of cultural monuments in the Middle East in recent years has instigated or given renewed energy to concerns about threatened or destroyed cultural heritage and the use of technology in restoration and preservation. Beginning with work of Armin Grün and his colleagues on the Bamiyan Buddhas, crowd-sourcing of photographs and the use of photogrammetry to construct 3D representations has become an important method of enhancing existing archaeological data in preservation and reconstruction of cultural monuments that are at risk or have been destroyed. Current work of this kind includes - but is by no means limited to - the Rekrei initiative of Matthew Vincent and Chance Coughenour, and the Center for Cyber-Archaeology and Sustainability directed by Thomas Levy, based at University of California San Diego but established as a consortium also involving UCLA, UC Berkeley and UC Merced.

The case of the Armenian cemetery at Julfa, near the ancient city of Jugha, is somewhat unique, because in the century prior to its final destruction access was very restricted, so archaeological research was not possible, and there are rather few – and certainly no 'tourist' - photographs to draw on. At the beginning of last century, there were 10,000 tombstones in the cemetery, including a very large number of ornately carved cross stones ('khachkars') that are unique to Armenian culture. The cemetery was completely destroyed by Azeri troops in 2005-2006, and the site converted into a military shooting range.

What we have as the basis for a reconstruction project are an archive of 2,000 photographs taken very systematically in the 1970s and 1980s by a remarkable Armenian scholar Argam Ayvanzyan, about 500 photographs taken on glass negatives in the early years of the 20th Century, and nearly 50 tombstones that were removed from the cemetery over the course of the century, providing us with valuable direct evidence.

The roots of Armenian culture can be traced to the establishment of Nakhichevan during the Fourth Century BC in what is now the Nakhchivan Autonomous Republic, an exclave of Azerbaijan. Nakhichevan's name derives from the Armenian "Nakhnakan Ichevan" (landing place), referring to the place Noah landed his Ark after the biblical deluge. It was in Nakichevan that Mesrob Mashtots first created the Armenian Alphabet and opened early Armenian schools. The centre of Nakichevan's culture was the ancient city of Julfa (or Jugha), destroyed by order of Shah Abbas in 1605 during one of the periodic wars between Persia and Turkey. The Shah's scorched earth policy did not require the destruction of the cemetery, so it survived.

Until 2005, Julfa cemetery graced the banks of the river Arax with 10,000 tombstones and other funerary monuments, including over 2,000 ornate Armenian khachkars (cross-stones) from the 15th and 16th century, inscribed with Christian crosses, suns, flowers and climbing plants. Alongside Julfa's khachkars stood heavily inscribed ram-shaped stones, unique to this cemetery, and ordinary tombstones. Spread over three hills on Nakhichevan's border with Iran, Julfa cemetery was home to the largest (and probably the earliest) collection of Eastern Christian cultural monuments in existence.

In 2005 Azerbaijani authorities demolished Julfa cemetery's priceless khachkars with bulldozers, loaded the crushed fragments onto trucks and emptied them into the river Arax. Shortly thereafter, Nakhichevan authorities constructed a military shooting range on the very ground where thousands of human remains lie, now unmarked.

In 2013 a small research team was established, seeking to ascertain whether sufficient primary sources still existed to make possible a large scale digital recreation of Julfa cemetery. The results were published in the ebook "Recovering a Lost Armenian Cemetery", which can be downloaded from the project webpage at (https://julfaproject.wordpress.com/). Based on this work, the Julfa Cemetery Digital Repatriation Project was launched in 2015. It aims to return to the Armenian people the entire medieval section of Julfa's cemetery, consisting of 2,000 khachkars and ram-shaped stones. These destroyed monuments are now designated by UNESCO as 'intangible world heritage'.

In further field work in 2015 and 2016, in Armenia, Iran and Georgia, the project team took over 50,000 high resolution digital photographs and 3D scans of the extant tombstones – i.e. those removed from the cemetery during the 20th Century. These were used to create an initial immersive 3D exhibition, which was shown for the first time in Rome in September 2016. Further shows are planned in Australia and North America by the time of the DH2017 conference, with additional khachkars added as the work on the photographic archive proceeds.

The Julfa Cemetery Digital Repatriation Project is hosted by Australian Catholic University at its North Sydney campus, and is a project of the Institute for Religion and Critical Inquiry. The purpose of the project is to create a virtual reconstruction of the cemetery, with the aims of ensuring the public memory of Armenia's cultural heritage, not only for the benefit of Armenians but also as an important contribution to world cultural heritage. The project will also restore a measure of dignity to the 10,000 deceased inhabitants, whose graves now lie unmarked beneath a shooting range. The project will also safeguard an important testimony to early Christian history in the Near East, and to Armenian-Persian and Christian-Islamic relations over a period of centuries.

The project's primary goals are six-fold:

- to create an extensive archive of materials related to the cemetery and its monuments–photographs, documents and digital materials, to be housed at Australian Catholic University and the State Library of New South Wales in Sydney;
- to carry out research and create a basis for ongoing research, not only in Armenian history, religion and culture, but in the history and culture of the wider region, including Persia/Iran;
- to create permanent virtual reality installations in Yerevan and Sydney (and any other city that wishes to have one);
- to create a touring exhibition that can travel to cities without the resources to establish a permanent installation;
- to create a vivid web presence, including online virtual reality exhibits, open to comment and contribution to everyone who may be interested;
- to work in collaboration with other projects and individuals interested in the preservation and reconstruction of destroyed and endangered cultural heritage.

The main source materials are over 2,500 original photographs of the site dating from 1915 until the present, illustrated manuscripts, handwritten journals, architectural sketches and audio recordings. The installations and other outputs will be derived from the repository of research materials gathered and created by the project—archaeological, historical, cultural, theological.

One distinguishing feature of the project is that in addition to ambitious 3D visualisations and realisations, research is being carried out on the carvings and inscriptions. The symbolism of the carvings is important in Armenian theology and cultural history. In addition, each khachkar was created for an individual (or occasionally an event), and the project is researching the inscriptions in order to identify as many as possible of the individuals (and events) they commemorate.

The reconstruction of the cemetery is of particular importance to Armenians, who see its destruction as part of a pattern of cultural genocide in the Near East, and the project has received considerable support from Armenians not only in the country but in the many diaspora communities around the world, including Sydney. There is much wider significance, however. The cemetery was on the border between Armenia and Persia, and the photographic record demonstrates considerable Persian / Muslim influences in the design and carving of the stones. The history of the cemetery is important, therefore, as Iranian as well as Armenian history. In addition there is the wider consideration that all cultural heritage is world heritage.

(Note: a key reason for the project being based at the Australian Catholic University in Sydney, Australia is that the Armenian diaspora in that city numbers 45,000, with many of the families living there able to trace their ancestry back to the Armenians forced by Shah Abbas to leave the ancient city of Julfa prior to its destruction, and to travel with his army to the city of New Julfa, which he constructed near his capital, Isfahan.)

One of the key challenges the project is facing is to do with the 'politics' of cultural memory. Whose interpretation of Armenian – or Iranian - cultural history should be represented? How can differing perspectives on the symbolism of the carvings be reflected? Engagement with a global 'audience' is a key commitment of the project, but how can such engagement be managed in a practical manner that respects the viewpoints and 'rights' of all who wish to contribute?

The political dimensions of repatriating conflict-destroyed sites and the 'human rights' aspects of the project are among some broader questions around the role of digital humanities scholarship in addressing cultural genocide and social injustice. How explicit should such motivations be, especially given that questions of cultural genocide are always contested?

Related to these questions, the paper will also consider how the Julfa Cemetery project stands in the landscape of the many other projects now at work on the reconstruction of threatened or destroyed cultural heritage, and the opportunities for collaboration, not only in relation to technical methods, but perhaps even more importantly in relation to the social, political and regional issues that are common to them all.

Bibliography

Ayvazyan, A (1993). *Khachkars of Djugha*. Yerevan: Hushardzan.

Ayvazgyan, A (2007). *The Symphony of the Destroyed: Jugha Khatchkars*. (Published in English, Russian, and Armenian.) Yerevan.

Bachoian, M (2011). "The Destruction of Djulfa." Law Journal for Social Justice. Arizona State University. April 6, 2011

Baltrušaitis, J and Kouymjian, D. (1986) "Julfa on the Arax and Its Funerary Monuments". *Armenian Studies*. In Memoriam Haig Berberian. Galouste Gulbenkian Foundation.

Crispin, McAlary, Riddell and Marshall (2014). *Recovering a Destroyed Armenian Cemetery: A pilot project*. Melbourne: Magnet Galleries.

Denard, H et al.(n.d.)*The London Charter for the Computer-based Visualisation of Cultural Heritage*. http://www.londoncharter.org

Galichian, R (2010). *The Invention of History: Azerbaijan, Armenia and the Showcasing of Imagination*. Gomidas Institute: London. Second Edition

Grün, A et al. “Photogrammetric Reconstruction of the Great Buddha of Bamiyan, Afghanistan.” *The Photogrammetric Record 19*(107): 177–199 (September 2004) - (available at http://www.researchgate.net/profile/Armin_Gruen/publication/227635047_Photogrammetric_Reconstruction_of_the_Great_Buddha_of_Bamiyan_Afghanistan/links/00b7d528dcf9161fd9000000.pdf)

Hasratian, M (1999). “L’art des khatchkars de l’école de Djougha.” *Environmental Design: Journal of the Islamic Environmental Design Research Centre* (in French)

High-Resolution Satellite Imagery and the Destruction of Cultural Artifacts in Nakhchivan, Azerbaijan (2015) (a report by the American Association for the Advancement of Science – Science and Human Rights Program) *https://www.aaas.org/page/high-resolution-satellite-imagery-and-destruction-cultural-artifacts-nakhchivan-azerbaijan.*

Marr, N (1983). “Des Monuments du cimetière de Djoulfa.” (In French.) *Xristjanskii Vostok*, vol. IV, no. 2, p, 198 et pl. IX

Montperreux, F D. de (1840). *Voyage autour du Caucase, chez les Tcherkesses et les Abkhases, en Colchide*. University of Lausanne (in French)

Research on Armenian Architecture (2006). “Julfa: The Annihilation of Armenian Cemetery by Nakhichevan’s Azerbaijani Authorities,” Report by Research on Armenian Architecture.

Petrosyan, H (2015). *Khachkar*. Zangak,

Petrosyan, H (2004). “Iconography of Jugha’s Cross-Stones.” Yerevan: Historical-Philological Journal, 2004, issue 1 (in Armenian with English and Russian summary)

The Julfa Cemetery Digital Repatriation Project (n.d.) *https://julfaproject.wordpress.com*

UC San Diego (n.d.) UC San Diego Center for Cyber-Archaeology and Sustainability *http://ccas.ucsd.edu*

Vincent, M., and Coughenour, C. (n.d.) Rekrei initiative: *https://projectmosul.org*

Vruyr, A (1915/1967). “Jugha”. Republished in Yerevan: Historical Philological Journal, 1967, issue 4 (in Armenian with Russian summary)

Yakobson, A (1978). “Historical-Artistic Observations about Armenian Khachkars.” Yerevan: Historical Philological Journal, 1978, issue 1 (in Russian with Armenian summary)

Can VR Survive Peer Review? Cultural Challenges for 3D Research

Lisa M. Snyder
lms@ats.ucla.edu
UC Los Angeles, United States of America

Alyson Gill
agill@umass.edu
UMass Amherst, United States of America

In the concluding days of the 2016 session of the NEH Advanced Topics in the Digital Humanities Summer Institute on Advanced Challenges in Theory and Practice in 3D Modeling of Cultural Heritage Sites held at UCLA, participants, faculty, and invited scholars focused discussion on the critical issues facing academics working with 3D content. The goal of these conversations were three-fold: 1) to clearly articulate the challenges facing researchers integrating 3D tools and methods into their scholarship, 2) to outline key questions and new lines of inquiry for future investigation, and 3) to develop actionable recommendations to position 3D work as a valid – and viable – mode of knowledge production. This paper describes that process, the topics chosen for discussion, and the resultant list of action items for the 3D community.

Hapax: Probabilistic part-of-speech tagging in XQuery and XForms

C. M. Sperberg-McQueen
cmsmcq@blackmesatech.com
Black Mesa Technologies LLC, United States of America

Many programs perform part-of-speech (POS) tagging on texts [Leech et al. 1983, Booth 1985, Church 1988, DeRose 1988, Brill 1992, Leech et al. 1994, Schmidt 1994, 1995, Toutanova et al. 2003]; although they use a variety of algorithms, their interfaces tend to be similar:

- They work in batch mode, not interactively.
- They generally model text as a flat sequence of characters; for most, XML markup must be removed before data are submitted to the tagger, and afterwards merged back into the output.
- They are consequently unable to exploit information in XML markup — for example, that “Brown” is here a proper noun and “Essex” there a place name.
- They tag each word token in the input with their best guess at the correct POS; by default, they do not distinguish low- and high-probability guesses.
- They cannot accept partially tagged input. In consequence, the human annotator cannot help them by providing hints on some words.
- They operate on words, not smaller segments.

This paper describes an XQuery-based POS tagger designed to differ from existing taggers in all of these ways. It works interactively one sentence at a time directly on XML (by default, TEI-encoded) text, exploits relevant markup, provides not just the most probable tagging of the input but several ranked alternatives, accepts partially tagged input, and can work on user-specified segments (e.g. TEI w [word] or m [morph] elements) instead of only on space-delimited tokens. Because the tagger described here is designed to support semi-automatic (or 'half-automatic') POS annotation for XML data, it has been given the name Hapax.

Hapax has been designed and implemented as part of the project "Annotated Turki Manuscripts from the Jarring Collection Online" (ATMO), supported by the Henry R. Luce Foundation; the author thanks both the Luce Foundation for their support and his colleagues in the ATMO project for their collaboration.

Design considerations

Early POS taggers used morphological and other rules to assign POS tags to input; later experience showed that purely statistical methods like hidden Markov models (HMMs) could achieve better accuracy with less effort; for tutorial descriptions of HMMs see Rabiner 1989 and Charniak 1993.

For batch-mode POS tagging, accuracy and speed are obvious desiderata. Many modifications, refinements, and alternatives to HMMs have been proposed; these can improve accuracy by several percentage points. Larger training sets make a much larger difference. Schmid 1994 reports a comparison in which the least and most accurate taggers differ by two to four percentage points, while accuracy rates for small and large training sets (< 10,000 and > 1,000,000 words) differ by twelve to sixteen points.

For Hapax, intended to support human annotators working on under-resourced languages, raw speed is unimportant. For any tagger, the human annotator will need to correct many proposed taggings; the key to improving annotation speed is to make corrections faster.

Selecting the correct tag from a menu requires several interactions: the 80 tags in the Brown Corpus POS tag set do not fit into a single menu; many tag sets are larger. Accepting a proposed tagging for a word requires a single user-interface interaction (e.g. clicking "OK").

So speed improves with accuracy: the fastest corrections are those not needed. But high accuracy requires large training sets, which under-resourced languages lack by definition. Some algorithms cope well with limited data. In the Brown Corpus, 92% of all tokens are tagged with the most frequent POS tag for their word type. A trivial 1-gram tagger, which just assigns the most frequent POS tag for each word form, will thus do almost as well on known words as more sophisticated algorithms. In reality, not all words are known, but a 1-gram tagger trained on as little as 2000 words from the Brown Corpus will tag 60 to 70% of input tokens correctly. Larger training sets (8000, 32000, 128000, 500000 words) again do better (68-78%, 73-85%, 77-90%, 82-92%).

Also, we can make tagging errors less costly to fix. If the tagger provides one tag for each segment, every wrong guess costs a manual tag selection. If the tagger proposes several POS tags, then some errors will be as cheap as a correct tagging: one mouse-click. So the goal of Hapax's design is to minimize the need to select tags from menus, by proposing not one but several POS tags for each word.

If a 1-gram tagger for the Brown Corpus proposes not one but three POS tags, the correct tag will be among those proposed 71-80%, 79-86%, 84-93%, 87-97%, or 90-98% of the time (for 2000-, 8000-, 32000-, 128000, 500000-word training sets). If five tags are proposed, the correct tag will be proposed 79-88%, 87-92%, 91-95%, 92-97%, or 94-98% of the time.

If a single user interaction can accept a proposed tagging for the entire sentence, we will save one interaction for each word of the sentence. Hapax uses a standard bigram HMM to calculate the N most likely taggings for the entire sentence. The higher N is set, the greater the chances that only a single mouseclick will be required, but more time will be needed for reading and considering the proposals; it is likely that there is a point of diminishing returns.

XQuery implementation

Hapax is implemented as a library of XQuery functions. One set of functions reads the training material and produces XML word- or POS-frequency lists from them. These list word types or POS tags by frequency, subdivided by POS tags or word types (or, for bigrams, POS of following segment). Additional functions calculate probability distributions for use with unknown words, using the technique of Charniak et al. 1993.

The 1-gram tagger consults the word/POS frequency list and returns the N most likely POS tags for the given word form. The bigram tagger consults the bigram and POS/word lists and uses the standard Viterbi algorithm to calculate the most likely path through the trellis of possible taggings for a sentence. A simple modification of the algorithm allows Hapax to calculate not one path but the best N paths, with time linear in the number of tags in the trellis.

Testing routines generate random test and training sets from a corpus stored as an XQuery database; in a project setting, the training sets are not created on the fly but prepared in advance and stored in a database.The primary interface for consumers of the Hapax library is the function hapax:tag(), which accepts as arguments:

- An XML element representing a sentence
- An indication of what frequency data to use
- Optionally, a set of access functions

The function calls the 1-gram and bigram taggers and returns an XML document describing possible POS taggings for the input. In the common case, the input sentence is a tei:s element, containing tei:w or tei:m elements to be tagged. Input elements may have type attributes; such a partial tagging of the sentence will affect the probabilities for the POS tags for other elements. The optional set of access functions allows Hapax to be used with non-TEI markup; the user-supplied functions are used to identify words in a sentence, detect POS tagging in the input, and add POS tags to the output.

The entire Hapax library is a few thousand lines of XQuery; the rich sets of data structures (including XML as a native type), higher-order functions, and grouping constructs in XQuery and XSLT make the implementation of POS-tagging algorithms remarkably straightforward.

XForms interface

In the ATMO project, Hapax supports a browser-based user interface specified with XForms. The form displays a document, providing an Annotate button for each sentence. When the button fires, the form sends the sentence to the Hapax back end and uses the response to build a form for accepting or changing the annotation. The most likely taggings for the sentence are shown, each with an Accept button. A "Tag word-by-word" button is also shown; in word-by-word annotation, each segment in the sentence is displayed with several proposed tags: first those in the full-sentence taggings, then other common tags for the word type, and a worst-case "Tag manually" button which exposes the POS menus. The user can tag one or more words and activate a "Re-annotate" button, which re-submits the sentence to the back end. This allows the user to explore the effect of one POS assignment on POS probabilities for nearby words.

Within the ATMO project, data must also be segmented and spelling-regularized; those topics and their interaction with POS tagging are not discussed here.

Further work

Hapax v1 uses standard 1- and 2-gram HMMs for POS tagging (Charniak et al. 1993). Future versions should implement Schmid's binary-decision-tree method (1994, 1995), which helps with sparse data. More challenging will be adapting the directed-graph model of Xuehelaiti et al. (2013) to probabilistic POS tagging. This two-level model would allow the probability of a given stem's POS tag to depend not only on the POS of the immediately preceding morpheme(s) but on the tag(s) of the preceding word stems, which may improve tagging accuracy for agglutinative languages.

Bibliography

Booth, B.M. (1985), "Revising CLAWS," *ICAME News* 9: 29-35.

Brill, E. (1992), "A simple rule-based part of speech tagger," in *Proceedings of the Third conference on applied natural language processing*, Trento 31 March - 3 April 1992 ([n.p.]: Association for Computational Linguistics), pp. 152-155.

Charniak, E. (1993), *Statistical language learning* (Cambridge: MIT Press).

Charniak, E., Hendrickson, C., Jacobson, N., and Parkowitz, M. (1993), "Equations for Part-of-Speech Tagging," in *Proceedings of the 11th National conference on artificial intelligence*, Washington DC July 11-15, 1993 ([n.p.]: The AAAI Press; Cambridge: MIT Press, 1993), pp. 784-789. Web. http://www.aaai.org/Papers/AAAI/1993AAAI93-117.pdf(Accessed: 1 October 2016)

Church, K. W. (1988), "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin 9-12 February 1988 ([n.p.]: Association for Computational Linguistics), pp. 136-143. Web. http://www.aclweb.org/anthology/A/A88/A88-1019.pdf. (Accessed: 1 October 2016)

DeRose, S.J. (1988), "Grammatical category disambiguation by statistical optimization," *Computational Linguistics* 14.1, pp. 31- 39.

Leech, G. Garside, R., and Bryant, M. (1994), "CLAWS4: The tagging of the British National Corpus," In *Proceedings of the 15th International conference on computational linguistics (COLING 94) Kyoto, Japan*, pp. 622-628. Web. http://ucrel.lancs.ac.uk/papers/coling.html . (Accessed: 1 October 2016)

Leech, G., Garside, R., and Atwell, E. (1983), "The automatic grammatical tagging of the LOB corpus," *ICAME News* 7: 13-33.

Rabiner, L. R. (1989) "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition."Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77.2: 257-286.

Schmid, H. (1995), "Improvements in part-of-speech tagging with an application to German," In *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland*. Revised version available on the Web at http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (Accessed: 1 October 2016.)

Schmid, H. (1994), "Probabilistic part-of-speech tagging using decision trees," In *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*. Web. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf . (Accessed: 1 October 2016)

Toutanova, K., Klein, D., Christopher Manning,, C. and Singer, Y. (2003), "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," In *Proceedings of HLT-NAACL 2003,* pp. 252-259.

Xuehelaiti, M., Liu, K., Jiang, W., and Yibulayin, T.. (2013). "Graphic Language Model for Agglutinative Languages: Uyghur as Study Case." *Chinese computational linguistics and natural language processing based on naturally annotated big data: 12th China National Conference, CCL 2013 and First International Symposium, NLP-NABD 2013*, Suzhou, China, October 10-12, 2013, Proceedings, ed. Maosong Sun. LNAI 8202. Berlin: Springer, pp. 268-279.

The Early History of Digital Humanities

Chris Alen Sula
csula@pratt.edu
School of Information, Pratt Institute
United States of America

Heather Hill
hhill4@pratt.edu
School of Information, Pratt Institute
United States of America

Most commentators locate the origin of digital humanities (DH) in humanities computing of the mid-20th century. Dalbello (2011), for example, begins her account in 1946 with Roberto Busa's plans for the Index Thomisticus, a massive attempt to encode nearly 11 million words of Thomas Aquinas on IBM punch cards. This event (and the narrative that follows) is found throughout the literature, leading some to believe that early DH work "concentrated, perhaps somewhat narrowly, on text analysis (such as classification systems, mark-up, text encoding, and scholarly editing)" (Presner 2010, 6). Others seem convinced that DH is still only text analysis—and misguided in its approach (Fish 2012).

This paper presents an empirical perspective on the early history of digital humanities by tracing publications in two foundational journals (*Computers and the Humanities*, established in 1966, and *Literary and Linguistic Computing*, established in 1986), with particular emphasis on media types and authors' disciplines.

Background

Despite the variety and breadth of definitions of DH (e.g., Gold 2012 and Terras 2012), narratives of its history have been surprisingly homogenous. Hockey (2004) and later authors (Svensson 2009, 2010, 2012; Kirschenbaum 2010; Dalbello 2011) all ground DH in mid-20th century humanities computing, a view that is all but orthodox in short and anecdotal histories of the field. According to this narrative, DH begins in 1946 with the Index Thomisticus and proceeds through advances in corpus linguistics to the founding of the journal *Computers and the Humanities* (*CHum*) in 1966. These early projects are hindered by storage capacity, hardware costs, and processing limits; progress is slow. Though Svensson (2009) admits that not every article during this time is about text analysis, he notes that the field had narrowed enough by 1986 for *Literary and Linguistic Computing* (*LLC*) to supplant *CHum* as the premier humanities computing journal (note the journal titles). Hockey similarly describes the 1970s and 1980s as a period of "consolidation" of text analysis methods. As storage and processing capabilities increased from the late 1970s onward, structured electronic text and multimedia archives dominated the field, followed in the 1990s by Internet-enabled hypertexts, digital libraries, and collaborative editing. The overarching theme of this narrative is text, with the plot revolving around corpora of increasing size and susceptibility to machine analysis.

Though this account dominates historical views of the field, it raises four separate concerns. First, it privileges certain disciplines, projects, and tools at the expense of others (e.g., quantitative history, which is absent from the narrative). Second, it fails to chart an actual historical path from early work in text analysis to "big tent" DH (Jockers & Worthey 2011; Pannapacker 2011a, b), encompassing everything from digital archives and databases to GIS, network analysis, new publishing formats, digital pedagogy, and so on. Third, it precludes historicizing and contextualizing current work that falls outside of text analysis, which may lead to a lack of attention to method, its historical complexities, and points of convergence with related fields such as the social sciences. Finally, these histories all suffer from a lack of evidence; the narrative is assumed and applied rather than documented.

An alternative approach would attend to the various methods, platforms, and tools that animate current DH work and investigate their origins in the literature. Ball (2013), for instance, has drawn attention to longstanding interest in computers and technology within writing studies. Scheinfeldt (2014) has pointed to the historical importance of oral history in the history of DH. Significantly, Nyhan, Flinn, and Welsh's (2013) project, "Hidden Histories," collects and archives oral histories from those who worked in the field during its first decades. These efforts are in broad solidarity with the empirical history presented here and contribute to the growing number of heterodox histories of DH.

Methodology

The corpus for this study consisted of 1,334 research articles published in *Computing and the Humanities* (1966–2004) and *Literary & Linguistic Computing* (1986–2004). This end date reflects the final issue of *CHum* and predates wide circulation of *A Companion to Digital Humanities* (2003), which was important in shaping DH in many ways. We omitted introductions, reviews, conference reports, and other articles that did not primarily present original research.

We manually inspected each article for media type and applied one of six categories (e.g., text, image, sound, object, number, other). For articles that addressed more than one media type, we recorded 'multimedia'. After inspecting several hundred articles, we added 'technology' as a media type to accommodate articles primarily about technology (e.g., AI, databases, hardware), rather than its application to a particular media. Using all eight codes (see Table 1), we coded or re-coded all the articles.

Category	**Example article topics**
Text	markup for full-text publishing; database of the works of Pascal; use of software for teaching literature
Image	digitized map of land register of 1822; database of manuscript images
Sound	programs for reproduction of sounds; correcting errors in musical databases; analysis of pronunciation
Object	machine classification of features of archaeological objects; techniques for dating medieval inscriptions
Number	database of wages, money, and prices; report on statistics programs; using Mark IV to extract quantitative data from charters
Multimedia	digitizing *Beowulf* (manuscript images and text); recording live performance; video and speech generation
Technology	artificial intelligence; computer-assisted language learning; mainframe and microcomputer file formats; MS Word 3.0; PF474 string co-processor
Other	report on a center; advancements in publishing; reasoning with natural language

Table 1. Media types coded in this study

In addition to media type, we recorded information about each author's discipline(s) and country of institutional affiliation. In the case of faculty appointed to more than one department, we recorded the discipline as 'multiple,' reasoning that an interdisciplinary appointment is more than a simple conjunction of its constituent departments. For authors located outside of traditional academic departments, we used one of three codings, where appropriate: 'center', 'non-academic', or 'GLAM' (galleries, libraries, archives, and museums). The remaining cases were clustered into one of 21 broad disciplines spanning the humanities and other areas.

Finally, we recorded whether each article had a focus on teaching and learning (e.g., courseware, language learning software).

Data on media type, disciplines, and teaching and learning were visualized using the free software Tableau Public and are available online.

Findings

The number of articles published each year varies from 6–50 (see Fig 1), the latter owing mainly to a double issue of *CHum* published in 1994/1995, which we recorded as 1995 because of its copyright date. Given the varying number of articles per year, we report several figures below as relative percentages each year (relative to the total number of articles that year). In cases where the two journals are compared, we also report relative percentages (relative to all articles from that journal in the corpus) because there are nearly twice as many total articles from *CHum* as compared to *LLC*, given their years of coverage.

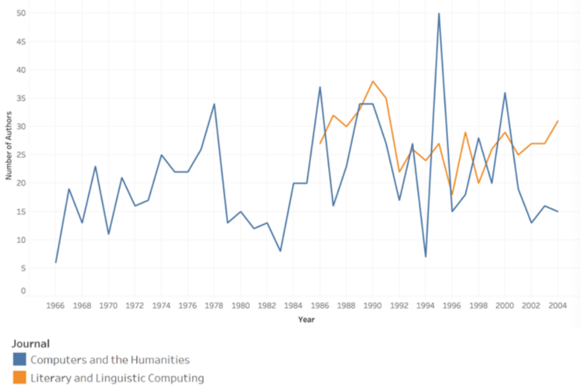


Figure 1. Number of articles published per year

Media type

Text is the most frequently studied medium (59% *CHum*, 72% *LLC*), but sound, multimedia, and reflections on technology are all present in the early literature (see Figs 2–3). These distributions vary by journal, with text being much more prominent in *LLC*.
'Other' is, admittedly, a rather large category at around 4% overall, but the heterogenous articles found there are not easily resolved into one or more media types, which is the focus of these codings, or even a primary theme, such as 'technology.' To some extent, many of these articles speak to the emergence of a field with its own meta-level discussions about theory and the production of knowledge. These articles are found throughout the early literature of DH and increase slightly around the end of the corpus, when "digital humanities" as such might be said to emerge.

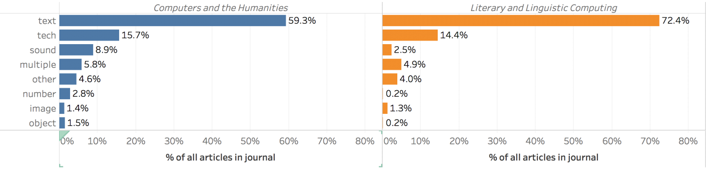


Figure 2. Articles by media type

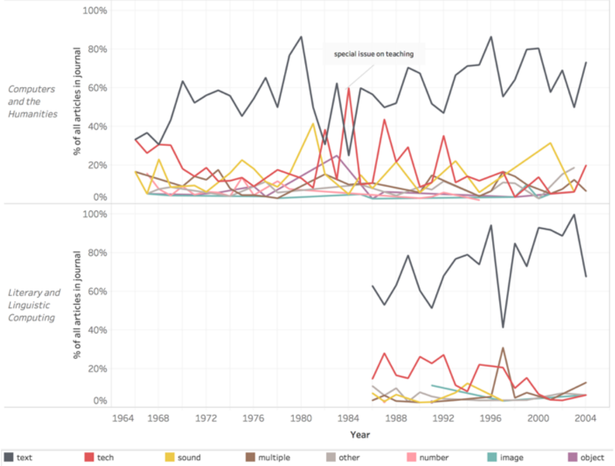


Figure 3. Articles by media type over time

Disciplinarity

The distribution of authors' disciplines present in each journal is shown in Fig. 4. Computing and computer science is most frequent, largely because of the amount of coauthors from those areas. English language and literature is the most frequent humanities disciplines, commensurate with Kirschenbaum's claim that DH's "professional apparatus…is probably more rooted in English than any other departmental home" (2010, 55). However, authors from languages and literatures departments other than English are nearly as common, as are centers, labs, and non-academic affiliations.

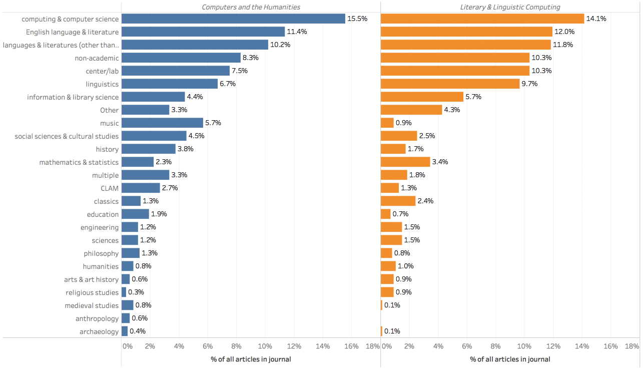


Figure 4. Articles by discipline

Some disciplines work with certain media types more than others (see Fig 5). For example, scholars of languages and literatures work almost exclusively with text, while art historians appear to favor multimedia.

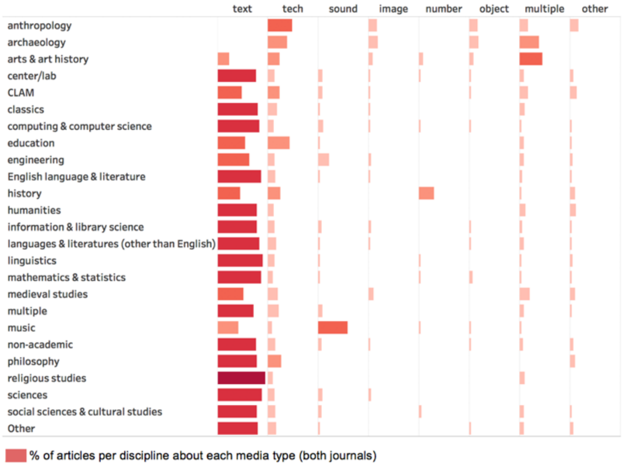


Figure 5. Media type by discipline

Location

Together, *CHum* and *LLC* represent nearly 50 different countries based on authors' institutional affiliations (see Fig 6).

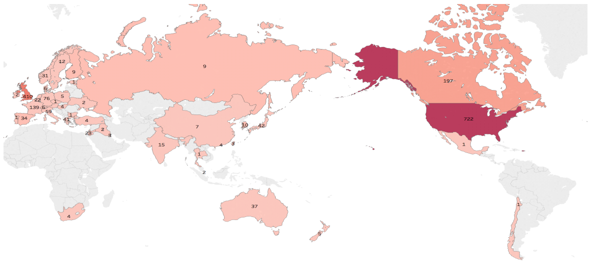


Figure 6. Location of authors

A small but appreciable portion of articles (5.6%, 75 articles) are international (i.e., with co-authors from institutions in different countries). However, the vast majority of authors in *CHum* and *LLC* hail from American and British institutions (respectively), though this predominance declines over the course of both journals (see Fig. 7). This data, as well as the largely Anglophone nature of these journals, presents a limited picture of early DH. A fuller analysis would include work published in other places and languages.

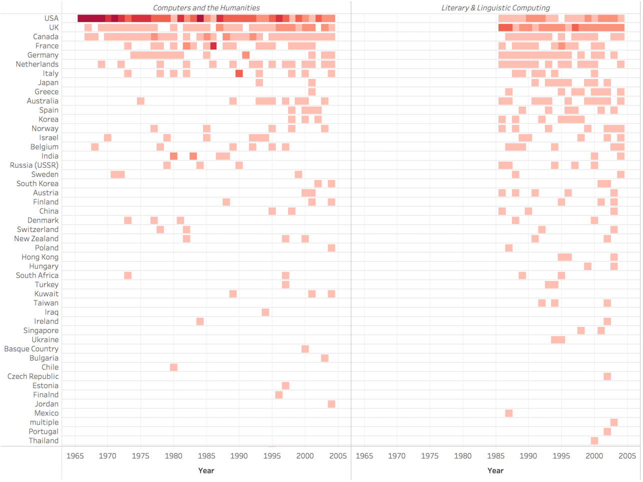


Figure 7. Location of authors over time

Teaching & Learning

There has been longstanding interest in teaching and learning in the field (as shown in Fig. 8), though less so within *LLC*. Peaks in each graph reflect special issues on teaching and learning published by each journal.

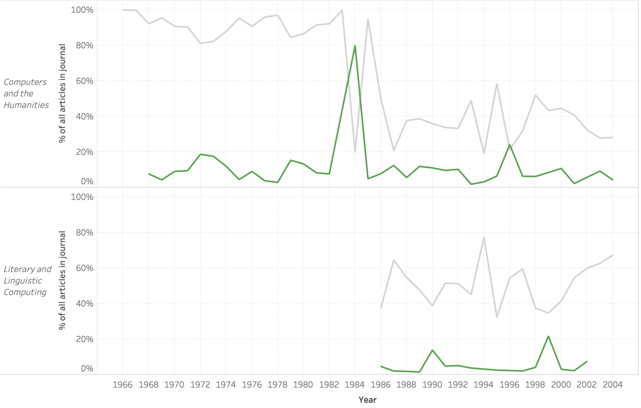


Figure 8. Articles about teaching and learning

Discussion and Future Directions

Rather than focusing on select disciplines, projects, tools, etc., this study includes the full range of early DH work (to the extent it appears in our corpus). The breadth of this picture helps set up the "big tent" view found in current accounts of the field. It also gives ground for historicizing and contextualizing the myriad forms of DH work today. One can imagine exploring this data to discover early DH articles about sound, in classics, from France, etc. and then consulting those primary source articles. Our study does provide some evidence for the claim that early DH work involves text experiments. Significantly, however, it documents the actual extent of that work (59% *CHum*, 72% *LLC*), and in so doing, highlights other work in the early history of the field.

Our next steps include exploring additional sources to expand our corpus. In part, this includes investigating disciplinary journals for early DH articles. We might also identify such articles or journals by mining citations in our *CHum*/*LLC* corpus or by consulting sources such as the

Companion. There are existing lists of early DH books as a starting point for monographs.

In addition, the full text of our corpus presents several possibilities for analysis, including a citation study that might address questions of transference between disciplines and the degree to which corpus articles cite each other (forming their own scholarly discourse) as compared to literature outside of core DH journals.

Bibliography

Ball, C. (2013). “Digital Publishing in the Tradition of Making Within Writing Studies.” *DH 2013.* Lincoln, Nebraska. http://prezi.com/8hrhbrqfw4fs/digital-publishing-in-the-tradition-of-making-within-writing-studies.

Dalbello, M. (2011). “A Genealogy of Digital Humanities.” *Journal of Documentation* 67 (3) (April 26): 480–506. doi:10.1108/00220411111124550.

Fish, S. (2012). “Mind Your P’s and B’s: The Digital Humanities and Interpretation.” New York Times, January 23, 2012. http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation.

Gold, M. K, ed. (2012). *Debates in the Digital Humanities.* Minneapolis: Univ Of Minnesota Press.

Hockey, S. (2004). “The History of Humanities Computing” in *A Companion to Digital Humanities.* Oxford: Blackwell, 1–19.http://onlinelibrary.wiley.com/doi/10.1002/9780470999875.ch1/summary.

Jockers, M. and Worthey, G. (2011). “Introduction: Welcome to the Big Tent.” *DH2011 Conference.* Stanford University. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-005.xml.

Kirschenbaum, M. G. (2010). “What Is Digital Humanities and What’s It Doing in English Departments?” *ADE Bulletin* 150: 55–61.

Nyhan, J., Flinn, A. D., Welsh, A. (2013). “Oral History and the Hidden Histories Project: Towards Histories of Computing in the Humanities.” *Literary and Linguistic Computing.*http://llc.oxfordjournals.org/content/early/2013/07/30/llc.fqt044.short.

Pannapaker, W. (2011a) “‘Big Tent Digital Humanities,’ a View From the Edge, Part 1,” *Chronicle of Higher Education,* 31 June 2011, http://chronicle.com/article/Big-Tent-Digital-Humanities/128434.

Pannapaker, W. (2011b) “‘Big Tent Digital Humanities,’ a View From the Edge, Part 2,” *Chronicle of Higher Education,* 18 Sept 2011, http://chronicle.com/article/Big-Tent-Digital-Humanities-a/129036.

Presner, T. (2010). “Digital Humanities 2.0: A Report on Knowledge.” http://cnx.org/content/m34246/1.6.

Scheinfeld, T. (2014). "The Dividends of Difference: Recognizing Digital Humanities' Diverse Family Tree/s" Found History. http://www.foundhistory.org/2014/04/07/the-dividends-of-difference-recognizing-digital-humanities-diverse-family-trees.

Svensson, P. (2009). “Humanities Computing as Digital Humanities” *Digital Humanities Quarterly* 3 (3). http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html.

Svensson, P. (2010). “The Landscape of Digital Humanities” *Digital Humanities Quarterly* 4 (1). http://www.digitalhumanities.org/dhq/vol/4/1/000080/000080.html.

Svensson, P. (2012). “Beyond the Big Tent.” In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 36–49. Minneapolis: Univ Of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/22.

Terras, M., Nyhan, J., and Vanhoutte, E. (2013). *Defining Digital Humanities: A Reader.* Ashgate Publishing, Ltd.

Humanités Numériques et Web Sémantique : du langage naturel à une représentation computationnelle structurée et sémantique des données

Pascaline Tchienehom
pkenfack@u-paris10.fr
Université de Paris 10, France

Résumé

ModRef est un projet du laboratoire Labex "Les passés dans le présent" qui accompagne divers projets sur des problématiques relatives aux humanités numériques (Oldman et al., 2014). Le projet ModRef s'intéresse spécifiquement au web sémantique (Berners-Lee et al., 2001) et aux données ouvertes et liées. Le but de ce projet est de réaliser une migration de données hétérogènes vers des triplestores encore appelés entrepôts ou collections de fichiers RDF afin d'améliorer le partage, l'échange et la découverte de nouvelles connaissances. Pour ce faire, la norme CIDOC-CRM (Boeuf et al., 2015) a été choisie car elle est actuellement la norme de référence pour la description sémantique de l'information muséographique ou d'héritage culturel (Hooland et Verborgh, 2014). Cette norme permet de décrire les caractéristiques globales des objets (identifiant, type, titre, matériau, dimension, note) mais également leur historique au travers d'évènements ou d'activités (origine, transfert de garde -localisations anciennes, localisation actuelle-, conservation) ainsi que les relations qui existent entre objets ou parties d'objets (bibliographie, composition, similarité, autre représentation -photo, dessin, tableau-, inscription). Par ailleurs, trois sous projets pilotes de ModRef ont été sélectionnés pour réaliser la migration des données : un conservatoire numérique de l'ensemble des documents rédigés en écriture cunéiforme, un corpus numérique d'objets archéologiques à iconographie mythologique et une bibliothèque numérique sur l'histoire de France du 20ième siècle. Les données de ces différents projets sont initialement décrites dans des bases de données ou dans des fichiers XML-EAD (Encoded Archival Description). Pour réaliser la preuve conceptuelle du projet ModRef, une architecture générale a été définie; une modélisation sémantique CIDOC-CRM et un alignement des données des différents sous projets pilotes ont été proposés; une migration des données vers des triplestores a également été effectuée. Une application web a été développée et déployée. Cette application web permet de décrire le projet ModRef mais aussi de consulter et d'interroger les triplestores créés. Les triplestores posent deux principaux défis scientifiques et techniques. Le premier est la migration de données souvent décrites initialement en langage naturel vers une représentation computationnelle structurée, puis sémantique de ces dernières. L'autre défi est l'exploitation des triplestores via des Endpoint Sparql (interface de saisie et d'éxécution de requêtes Sparql) ou via des interfaces sous forme de formulaires généraux d'interrogation.

Migration de données vers des triplestores

Une migration efficace et cohérente de données fait appel à différentes compétences. Pour assurer la pérénisation de cette procédure, une architecture générale et rigoureuse du workflow des différents types de données à manipuler doit être définie. Cette architecture explicite la démarche globale de tout projet qui souhaite faire migrer ses données vers des triplestores. Cette démarche se subdivise en différentes étapes bien identifiées : préparation des données (étude et description structurelle), modélisation sémantique et alignement des données structurées avec le modèle sémantique et enfin création et exposition des triplestores qui vont alors pouvoir être consultés et intérrogés. Notons que initialement les données sont souvent non structurées ou semi-structurées (notes, rapports, livres, html) et qu'il faut dans un premier temps en extraire une représentation structurée (tableurs, base de données, fichiers XML) pour pouvoir ensuite construire leur représentation sémantique plus facilement. Ce continuum d'étapes fait intervenir des compétences diverses et nécessite parfois d'adjoindre des profils intermédiaires entre deux étapes pour assurer le passage d'un format de représentation de données à un autre : (1) données non structurées ou semi structurées vers données structurées, et (2) données structurées vers données sémantiques. Par ailleurs, l'élément clé de l'architecture de la migration de données vers des triplestores est la modélisation et l'alignement des données avec le modèle de graphe sémantique choisi. Un graphe sémantique est un ensemble de noeuds et d'arcs orientés qui obéissent à un certain nombre de contraintes et règles (raccourci, héritage, inverse, symétrie, transitivité). Ce sont ces règles et contraintes qui définissent la cohérence et la validité d'un modèle. Nous avons utilisé la version 6.2 de mai 2015 du CIDOC-CRM qui définit 94 classes et 168 propriétés ainsi que son implémentation par l'Université d'Erlangen-Nuremberg. Afin de réaliser la migration, il a fallu procéder à un alignement des données avec certains noeuds du graphe sémantique à partir des informations extraites de bases de données ou de collections de fichiers XML-EAD. Les noeuds remplis par des valeurs sont des noeuds terminaux et les noeuds intermédiaires sont remplis avec des URIs qui définissent ainsi des chemins vers les noeuds terminaux. Notons qu'une rigueur particulière doit être apportée à la construction des URIs, à la fois pour leur lisibilité mais également pour la cohérence des chemins dans le graphe afin d'éviter des conflits de chemins et garantir ainsi l'unicité d'un chemin donné par rapport à un autre. Nous avons identifié les classes utiles (menant vers au moins une valeur non vide) pour modéliser les données des projets pilotes. Ainsi, la modélisation et l'alignement effectués représentent des extraits de graphes relatifs aux quatre thèmes suivants : (1) caractéristiques générales (identifiant, type, titre, matériau, dimension, note), bibliographie, composition et similarité d'objets; (2) évènements de début d'existence (origine) et de fin d'existence; (3) activités diverses (transfert de garde, conservation, mesure); (4) inscriptions et autres représentations (photo, dessin, tableau). De façon générale, ces extraits sont assez stables pour tout projet car, dans le CIDOC-CRM, il est possible d'identifier les chemins possibles menant à une information donnée sur un objet. L'alignement n'est pas une tâche programmatique mais fait appel à des détails de structure logique propre au modèle de description de données choisi par chaque sous projet. C'est une tâche à mi-chemin entre la modélisation et l'implémentation qu'elle permet d'entrevoir un peu plus clairement. L'alignement définit ce à quoi correspond chaque noeud de notre graphe et il ne reste plus qu'à générer les fichiers CIDOC-CRM correspondants tout en respectant la syntaxe de la norme RDF. Les triplestores créés vont ensuite être exposés pour consultation (sous trois formes : *rdf, triplets et résumé attribut-valeur*) et interrogation (*formulaires généraux et Endpoint Sparql*) via notre application web. L'exploitation des triplestores via l'interrogation et l'exploration de ces derniers et les bénéfices que l'on peut en tirer est l'autre aspect majeur autour de la question de ces nouveaux entrepôts de documents RDF que sont les triplestores.

Exploitation des triplestores

L'intérêt des triplestores est qu'on a un modèle connu public et publié de représentation de l'information ce qui permet d'interroger les triplestores indifféremment avec des procédures identiques. Nous avons défini deux procédures d'exploitation de nos triplestores : des interfaces sous forme de *formulaires généraux* et des *Endpoint Sparql*. Les formulaires généraux sont un moyen simple et assez intuitif, car très proche du langage naturel, pour formuler des requêtes vers nos triplestores. Il suffit de remplir les rubriques du formulaire qui nous intéressent et de lancer la recherche. Une requête Sparql est automatiquement construite à partir des valeurs des champs renseignés du formulaire et c'est cette requête qui est utilisée pour interroger le triplestore. Au terme de l'exécution de la requête, une liste d'objets séléctionnés est renvoyée en résultat à l'usager. Par ailleurs, on peut aussi interroger nos triplestores via des Endpoint Sparql. Ce deuxième mode d'interrogation nécessite la connaissance du langage Sparql qui est aujourd'hui le langage de référence pour l'interrogation de documents RDF. Sparql est un langage assez simple mais pas toujours à la portée de tous. Ainsi, les formulaires généraux peuvent être vus comme un premier point d'entrée pour l'interrogation des triplestores tandis que les Endpoint Sparql assurent une exploitation plus large de ces triplestores via une formulation libre de requêtes de type "Select". Notons que la notion d'exploitation de triplestores fait appel aux notions d'interrogation et d'exploration de graphe. Ainsi, l'interrogation de triplestores consiste à formuler une requête Sparql pré-formatée (formulaires généraux) ou libre (Endpoint Sparql) tandis que l'exploration de triplestores est une forme d'interrogation uniquement possible via des Endpoint Sparql qui permet aussi de découvrir différents chemins dans un graphe sémantique vers des données précises. En effet, plusieurs chemins peuvent permettre d'obtenir une même information dans un graphe (usage de diverses notions : raccourci, héritage, inverse, raffinement), sachant que ces chemins ne sont pas toujours tous renseignés. On peut donc écrire des requêtes Sparql pour découvrir si différents chemins vers une donnée précise existent ou pour connaitre les noeuds terminaux. L'exploration est donc importante pour s'approprier un triplestore spécifique. L'exploration permet aussi la comparaison de différents triplestores qui décrivent des données similaires (objets d'une même période historique, objets de même type, objets identiques) dans un contexte de LOD (Linked Open Data), par exemple. Ainsi, la comparaison de chemins assure une meilleure découverte des connaissances et augmente la correction ou l'enrichissement mutuel des connaissances des différents acteurs du LOD. Notre application web fournit un LOD pour ModRef ainsi qu'une liste de modèles de requêtes Sparql pour interroger, explorer et valider nos triplestores séparément ou ensemble. A plus long terme, l'objectif est d'intégrer d'autres LOD sur internet (Beek et al., 2016) (Daga et al., 2016) pour un partage, un échange et une découverte de nouvelles connaissances à plus grande échelle. Ainsi, le LOD doit améliorer la découverte de nouvelles connaissances, du fait de l'usage de formalismes, de langages de métadonnées, de thésaurus publiés, standardisés voire normalisés.

Remerciements

L'auteur remercie le laboratoire Labex "Les passés dans le présent" de l'Université de Paris 10 et le projet ANR ModRef de référence ANR-11-LABX-0026-01.

Bibliographie

Beek, W., Rietveld, L., Schlobach, S., et van Harmelen, F. (2016). Lod laundromat: Why the semantic web needs centralization (even if we don't like it). *IEEE Internet Computing* 20(2), 78–81.

Berners-Lee, T., Hendler,, J., et Lassil, O. (2001). The semantic web. *Scientific American*.

Boeuf, P. L., Doerr, M., Ore, C.E., et Stead, S. (2015). Definition of the cidoc conceptual reference model, version 6.2. *Produced by the ICOM/CIDOC Documentation Standards Group, Continued by the CIDOC CRM Special Interest Group*.

Daga, E., d'Aquin, M., Adamou, A., et Brown, S., (2016). The open university linked data data.open.ac.uk. semantic web. *Semantic Web* 7(2), 183–191.

Hooland, S.V., et Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums. How to Clean, Link and Publish Your Matadata*. ISBN 978-0-8389-1251-5.

Oldman, D., Doerr, M, , de Jong, G., Norton, B., et Wikman, T., (2014). Realizing lessons of the last 20 years : A manifesto for data provisioning and aggregation services for the digital humanities. *D-Lib Magazine* 20(7/8).

Replication, Visualization & Tactility: Towards a Deeper Involvement of 3D Printing in Humanities Scholarship and Research

Aaron Tucker
atucker@ryerson.ca
Ryerson University, Canada

In a chapter that begins with the question "What would you make it you had a machine that could make anything?" Lipson and Kurman rhapsodically illustrate the myriad uses for 3D printing, from "3D prints [of] a precise, highly detailed replica of [a] fetus" to "the not so distant future [where] people will print 3D living tissue, nutritionally calibrated foods, and ready-made, fully assembled electronic components" (7). On-campus maker spaces often tout 3D printers with the same utopian vigor, yet there is still a great deal of opacity around how introducing 3D printed objects into a Humanities classroom or scholarship would add value to those spaces or work. In my experience collaborating and working in the Digital Media Experience Lab at Ryerson University, when asked to speak with colleagues or students about the potentials for projects that involve 3D printing, two main types of anxiety arise: first, there is the intimidation that comes with learning a new set of hardware and software; and second, there is a large amount of trepidation around what to the technology should even be used for. While the first set of fears can be somewhat mitigated by the sort of community-friendly maker space environment that 3D printers are often housed in, the second set of concerns asks questions that are at the root of not just 3D printing in the Humanities, but Digital Humanities (DH) projects as a whole:

- What are the elements of 3D printing that make it a unique contribution to the Humanities? What are the technology's strengths and how might they best be harnessed?
- What are the limits of such a technology, from both the hardware and software perspectives?

- As a physical object imbued with potential meaning, how might an tactile object speak to issues of critical theory
- What are the components of 3D printing that lend itself to a powerful learning or scholarly environment?
- What can 3D printing do that other modes of interface, data visualization cannot?

Scholars like Mark Stefik explains that "digital sensemaking" most often takes place in the ecosystems of "digital information infrastructure, such as today's web and search engines" and, as such, a great number of projects focus on a DH understanding of digital sensemaking centre around the digital components of virtual object creation. Yet, as Ian Foster argues, "Informational technology can also enhance our abilities to make sense of information, for example, by allowing exploration via visual metaphors" (19). 3D printed objects have the unique ability to make physical/concrete, at varying scales with easy replication, abstract ideas and give scholars and students modes of engagement with metaphors and issues and are not present in other mediums (print, film, music, etc). Drawing from theory and history around sculpture, this paper will ask anyone wanting to begin a 3D printed project, **what** the object they wish to create a metaphor for is, as a physical object imbued with potential meaning. If a student or scholar wants to speak to certain issues, what objects might be the best metaphors for the arguments and issues they wish to discuss? What does adding the layer of technology in 3D printing then add to that argument? What do the specific properties of replication, visualization and tactility add to the metaphor and the argument being made? 3D Printing then becomes another mode to explore visual metaphors but with the unique and obvious understanding that such an object is immediately and equally digital and analog.

Rooted in Jentery Sayer's work in discussing 3D printing alongside Lipson and Kurman's "10 principles of 3D printing" (20-24), this paper will begin to answer the above questions by outlining three core considerations, with examples, in an attempt to foster further discussion about how a relatively nascent popular technology might best be understood and undertaken in a Digital Humanities project or classroom.

First, the nature of 3D printing is built around the notion of replication: once virtual objects are constructed, they can be repeatedly printed quickly and easily. This is a massive strength when considering a large scale multitudinous project, or a classroom environment wherein students may be asked to each design or find and print an object. When printing with recyclable material, the technology lends itself exceptionally well to iteration, encouraging prototyping and low-risk failure in service of a finished product. However, the notion of replication also extends to the technology's ability to mirror "real life" (and future, fantasy) objects: projects like Morehshin Allahyari's Material Speculation are able to recreate lost objects destroyed by ISIS and reprint them, effectively regenerating a version of the physical object that both evokes the original and challenges its audience by layering the technological on top of the original craftsmanship/artistry. I will blend this with discussion of Odile Fillod's project printing models of the clitoris as an educational and feminist tool of inquiry.

Second, 3D printing allows creative and untapped modes of visualizing data. A number of data visualization tools, especially for beginners in DH, are going to be relatively simple 2D graphics, such as charts, graphs, maps etc; advanced tools will include movement, interactivity, and pleasant aesthetics. Yet, the interfaces that 3D printed objects promote are distinctive in their mix of the physical and virtual. Using examples from Lipson and Kurman's chapter "A Factory in the Classroom," as well as my own work with translating poems into small landscapes-type pieces via a height map application as part of Loss Sets, it is clear that 3D printing offers a wealth of opportunities to translate numbers, words, spaces into objects; in doing so, the object is pushed into an analog space that challenges its audience to "read" and "understand" that data in an embodied and physical manner that maintains the pleasingly interactive and evocative arguments about the information selected often found in virtual data visualization tools.

Lastly, 3D printed objects have a sensual tactility that is difficult to get from other (virtual) elements of Humanities scholarship. Once printed, the objects have concrete weight and volume that points to a set of aesthetic values again straddling their digital-analog nature; they can be picked up, turned over, and explored from a multitude of angles. Too, as wearable computing increases, 3D printed fabric projects an engaged scholarship linked immediately to the sensations and shapes of the body. Generating projects that take advantage of the physicality and materiality of such objects can be an exceptionally effective and emotive mode to considering virtual or past objects, especially as printed objects can be treated, post-printing, to change the plastic's original properties. To this end I will be looking closely at Neri Oxman and the Mediated Matter Group's Lazurus and Vespers, a "series of death masks" as well as Donna Szoke's Decoy.

Before considering potential software or materials and hardware, a DH project involving 3D printing will be most effective if it begins by considering these principles, understanding the strengths and core of the technology itself, and then blending each element together into a symbiotic environment in which the object itself is capable of housing the necessary evocative complexity and wonder that the technology itself often provokes.

Bibliography

Allahyari, M. (n.d.) Material Speculation: ISIS. 3D Printed Objects, Resin filament. http://www.morehshin.com/material-speculation-isis/

Crompton, C., Lane, R.J., and Siemens, R.G (eds.) (2016) Doing Digital Humanities: Practice, Training, Research. Routledge.

Devon, E., MacDougall, R., & Turkel, W.J. (2012). "New Old Things: Fabrication, Physical Computing, and Experiment in Historical Practice." Canadian Journal of Communication [Online], 37.1: n. pag. Web. 19 Mar. 2017

Foster, I. (2011). "We digital sensemakers." Switching Codes: Thinking through Digital Technology in the Humanities and the Arts. Ed. Bartscherer, Thomas, and Roderick Coover. University of Chicago Press.

Lipson, H., and Kurman, M. (2013) Fabricated: The New World of 3D Printing, John Wiley & Sons,.

Sayers, J., Elliott, D., Kraus, K., Nowviskie, B. and Turkel, W. J. (2015) Between Bits and Atoms, in A New Companion to Digital Humanities (eds S. Schreibman, R. Siemens and J. Unsworth), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781118680605.ch1

Sayers, J. (2015a). Prototyping the past. Visible Language, 49(3), 157-177.

Sayers, J. (2015b) "Why Fabricate?." Scholarly and Research Communication [Online], 6.3: n. pag. Web. 19 Mar. 2017

Stefik, M. (2011). "Scholarsource: a digital infrastructure for the humanities" Switching Codes: Thinking through Digital Technology in the Humanities and the Arts. Ed. Bartscherer, Thomas, and Roderick Coover. University of Chicago Press.

Szoke, D. (n.d.). Decoy. 3D Printed Objects, PLA Filament. http://donnaszoke.com/?projects=cloud.

Tucker, A. et al. (n.d.) Loss Sets. 3D Printed Objects, PLA and ABS Filament. http://aarontucker.ca/3-d-poems/

Berry, D.M. (ed). 2012. Understanding Digital Humanities. Palgrave Macmillan, 2012.

Towards Feminist Data Production: A Case Study From Comics

Alexander Turton
a.turton@uea.ac.uk
University of East Anglia, United Kingdom

Introduction

This paper will take my ongoing doctoral research designing resources for analysing individual graphic novels as a case study and starting point to discuss how we can produce data in the Humanities which can be called 'feminist'. It will engage with the significant debates on this topic which are emerging in the Digital Humanities (Clement, 2016) (Drucker, 2012) (Losh et al., 2016) (Posner, 2016) (Rhody, 2016), particularly the panel discussion on creating feminist infrastructure at DH 2016 (Brown et al., 2016). Marking up images and designing ways in which to make them communicate with a comic's words are highly subjective enterprises. By explaining how I dealt with this issue in my research, this paper will outline how the subjectivity involved in creating our own data structures and ontologies, and these elements' inherent statuses as arguments about data, is a strength, an affordance, of such an approach, as well as something that embraces the situatedness and plurality of data that feminists in the sciences (Haraway, 1988) (Irigaray, 1985) have advocated for and which, more recently, have been advocated for in the Digital Humanities. Although my case study focuses on contemporary comics, this paper will explain how some of the resulting principles can be employed by data creators today in other disciplines and the GLAM sector.

Background and method

As a Digital Humanist operating outside of a specific department or centre, the question that I hear most often about my work is how I make the resources which I create objective, how I avoid my datasets merely reflecting one individual's interpretation of a text. But when I mark up an image in a certain way, or use a database structure to reflect the rhizomatic structure of a graphic novel, I do so not because it is 'appropriate' or a 'good fit', but because that is an argument I want to make about comics and their meaning mechanism, and by applying that algorithm to the dataset that is the comic, I articulate that argument, mobilising it and making it available for evaluation. I am not trying to enact as little violence as possible to a text; I am making an argument about it. This idea that datasets, data structures and algorithms are arguments that are made about texts or other objects of study is relatively well-established in Digital Humanities (Ramsay and Rockwell, 2012) but it is more often framed as a caution than an opportunity.

Working on contemporary comics, where there is no pre-existing database, and no automated or straightforward tagging of images, it would be easy to see mark-up or tagging as a hurdle, and a problematic one at that, given the fraught nature of remediating pictorial information into values that can be entered into a database. But, although I must design my own data tags and ontology, I do not have a mandate to preserve, gatekeep, or distribute otherwise inaccessible data since my objects of study are widely available. Focussing on individual texts, too, affords me time to spend designing tailored data tags and ontologies. I do not need to preserve, that is my freedom; I cannot be 'objective', that is my strength. My objects of study are relatively small; that means I do not have to be singular, I can be multiple.

Digital approaches, especially to comics (Walsh, 2012) (Dunst et al., 2016), often rely on a single categorisation of each entity or attribute. This paper will argue, rather, that our databases ought to be multiple. Rather than text-mining, a metaphor which suggests the removal of gratuitous material, I would encourage thinking of this practice as data curation, or rather, curations. Consider the analogy of a virtual museum with access to a complete catalogue of material – for is this not like our complete texts? – where anybody can hang the material in whatever way, in whatever ways, they choose. Different paths through the information can be curated, different logics created, retaining the plurality of signification that each piece holds, resisting positivism. This may well tell us as much about our hypothetical hangers as about the hung objects, but therein lies an opportunity. If curation – like datasets, data structures and algorithms – is argument, then why not bring multiple perspectives into conversation? And if we can represent data multiplicitously, we can do it investigatively. By creating multiple ontologies and data tags it is possible to embrace Brian Massumi's judgement, "[t]he question is not: is it true? But: does it work? What new thoughts does it make possible to think?" (Deleuze and Guattari, 1987: xv)

Since datasets are arguments, marking up the same data differently allows researchers to avoid asserting single values for complex, or simply ambiguous, pieces of information. It also liberates us to encode arguments we disagree with, or at least that we concede are problematic, in order to better understand or critique them. The reduction of gender to a binary, for instance, has been highlighted as an issue in quantitative approaches (Clement, 2016). This, of course, drags up familiar tensions between anti-essentialism and feminism, but if we can encode different modes of representing gender – or any other 'attribute' which is better represented on a spectrum – including the reductive binary mode, we can maintain the plurality of our data, whilst retaining the possibility to see how the text subverts such a binary categorisation; we bring the text to bear on the theory and in so doing, better understand the theoretical position of the text. Image tagging can operate in a similar way, by tagging the same pictorial signifier in multiple ways we can tag with an intention to investigate, not merely minimise violence to the text. By contrasting and combining different ontologies it is possible to shed light on our texts and to allow our texts to shed light on our ontologies, all the while fracturing any notion of computational methods as objective black boxes by foregrounding their artificiality.

Conclusion

Embracing a conceptualisation of Humanities data as complex and plural, this paper will use examples from my own research remediating graphic novels into databases to demonstrate how deploying multiple tags and multiple ontologies not only instantiates a more feminist approach to data but is actually a productive methodology for analysing texts. It champions not the analysis of datasets, but rather an analysis by datasets. As Laura Mandell said in Krakow, we need "metadata built for thinking, not sustainability." (Brown et al., 2016)

Bibliography

Brown S., Clement, T., Mandell, L., Verhoeven, D., Wernimont, J. (2016). "Creating Feminist Infrastructure in the Digital Humanities." *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 47 - 50.

Clement, T. (2016). "Where is Methodology in Digital Humanities?." In Gold, M. and Klein, L. (eds), *Debates in the Digital Humanities*. University of Minnesota Press, pp. 153 - 75.

Deleuze, G. and Guattari, F. (1987). *A Thousand Plateaus*. trans. by Brian Massumi. London: Continuum.

Drucker, J. (2012). "Humanistic Theory and Digital Scholarship." In Gold, M. (ed), *Debates in the Digital Humanities*. University of Minnesota Press, pp. 85 - 95.

Dunst A., Hartel, R., Hohenstein, S., Laubrock, J. (2016). "Corpus Analyses of Multimodal Narrative: The Example of Graphic Novels." *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 178 - 180.

Haraway, D. (1988). "The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies*, 14(3): 575 – 99.

Irigaray, L. (1985). "Is the Subject of Science Sexed?" trans. by Oberle, E. *Cultural Critique*, 1: 73 – 88.

Losh, E., Wernimont, J., Wexler, L. and Wu, H. (2016). "Putting the Human Back into the Digital Humanities: Feminism, Generosity, and Mess." In Gold, M. and Klein, L. (eds), *Debates in the Digital Humanities*. University of Minnesota Press, pp. 92 - 103.

Posner, M. (2015). "The Radical Potential of the Digital Humanities: The Most Challenging Computing Problem is the Interrogation of Power." *Impact of Social Sciences Blog*. Web. 15 Sept. 2016.

Ramsay, S. and Rockwell, G. (2012). "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities." In Gold, M. (ed), *Debates in the Digital Humanities*. University of Minnesota Press, pp. 75 - 84.

Rhody, L. (2016). "Why I Dig: Feminist Approaches to Text Analysis." In Gold, M. and Klein, L. (eds), *Debates in the Digital Humanities*. University of Minnesota Press, pp. 536 – 39.

Walsh, J. (2012). "Comic Book Markup Language: An Introduction and Rationale." *Digital Humanities Quarterly*, 6(1).

Iterative Data Modelling: from Teaching Practice to Research Method

Pim van Bree
pim@lab1100.com
LAB1100, The Netherlands

Geert Kessels
geert@lab1100.com
LAB1100, The Netherlands

Introduction

Data modelling is an essential process of almost any digital humanities project (Flanders and Jannidis, 2015). Whether texts, images, or any other form of data is mapped or analysed, a model has to be conceptualised that describes the data and forms the bedrock of the application that contains or analyses the data.

Since data modelling in the humanities is largely perceived as an epistemological process, rather than an ontological process, there is a tension between the way in which material and knowledge presents itself and the manner in which material and knowledge can be described on a generalised or abstracted level. As Flanders and Jannidis (2015: 236) have pointed out: "Some of the most fertile and urgent areas of digital humanities research involve the question of how to develop data modeling approaches that accommodate both the self-reflexivity required by humanities research and the actionability and computational clarity required by the digital domain."

In this paper we reflect on a data modelling approach that has proven to be an effective teaching practice as well as a useful research method. The iterative data modelling approach we put forward focuses on a continuous shift between three levels of data modelling: conceptual level, logical level, and interface level. We have found that this approach provides students and scholars in the humanities ("scholars") with the skills they need to translate their body of data or research question into an operational process that produces rich (inclusive; fuzzy and uncertain) and complex (advocate divergent classes) actionable datasets. It is important to note that even though it is useful when a scholar can develop their own data model for computer-aided analytical purposes, we should not underestimate the learning curve this new skillset requires.

This paper focuses on experiences we gained from data modelling practices in the humanities aimed at developing a relational database. We draw on the results of over 20 courses and workshops for scholars we have held in the past three years on developing data models and using database applications. These insights are also informed by the continuous development of the research environment nodegoat, developed by the authors of this paper, and the scholarly collaborations resulting therefrom.

Challenges

Most scholars do not perceive their material or knowledge as 'data' (Posner, 2015). Once a scholar has accepted that lists of people, statements, and ideas can also be seen as data and that we do not necessarily need to be able to count with them, it becomes clear that their material or knowledge can be modelled as well. Like the analog card catalog, a database helps to store data properly and sustainably. This allows us to filter and query the data. We can then also create networks and analyse relationships. It is important to note that vagueness, uncertainty, and incompleteness can be incorporated in a data model.

To allow scholars to operationalise the data modelling process, three levels of a data model have to be studied. The conceptual level, the logical level, and the interface level describe the data at hand, each in its own way. Here, it is necessary to reflect critically on hidden assumptions in the choice of entity types and classifications (Erickson, 2013). An iterative data modelling approach is largely research driven, although existing standards could be used as well. By asking scholars to operationalise their own data model, rather than using or implementing a pre-existing model, they get acquainted with the complexities and granularities that operationalising a data model entails.

The interface challenge - how to operate a database application? - is very important. We see the translation of the data model into an actual database as a vital step to get a good understanding of the data modelling process. We prefer to do this with a database application that has a graphic user interface to be able to iterate quickly and to easily compare data models.

Teaching Practice

The participants in our data modelling workshops ranged from undergraduates to established scholars. These workshops were either in the format of intensive one day workshops or stretched over multiple events in the course of months. During a workshop, we first addressed the aforementioned challenges to show that the challenges participants face are not new and that we can critically reflect on them. Secondly, we did collective exercises to give participants an understanding on how data models and database applications work.

When we then asked them to conceptualise a data model based on their own research question, most participants did not know where to start. The reason for this seemed to be twofold. First, they were unable to process new information regarding data modelling and the database application into an operational process. Since most of the participants were trained to conduct research with a syntagmatic dimension in mind, a linear text, it was hard to execute a research process that leads to a paradigmatic dimension, a database (Manovich, 1999). Secondly, since they were invested in the complex and unique aspects of their research project, they were unable to operationalise a coherent model while keeping relevant variables and complexities in mind (Beretta, 2016).

To overcome this, we introduced an iterative approach that took them back and forth from their research question to a partial conceptual data model, to a partial logical model, and to a partial functional database application. This process helped them to first understand how to translate one class of information to a single, non-relational, data table. Once they could process basic typed values (strings/numbers), they started to work with texts, images, dates, and locations. These steps informed participants on the transformation of a conceptual idea into a table with fields in a database application. They first focused on the basics, the finite, while leaving growing complexity, the infinite, to next iterations: creating additional data tables and constructing relationships between them. After these practical questions had been tackled, attention was shifted towards uncertain data, fuzzy data, and the question on using existing standards for a data model.

In literature on data modelling processes, a distinction is made between the conceptual/logical level and the level of the application. A data model should be portable and not dependant on one application (Flanders and Jannidis, 2015). However, this does not mean that the conceptual/logical level may not be informed by the application while teaching data modelling practices. The feedback loop between these different levels has proven to be an essential step in helping scholars understand how their own research project can be translated into a data model and a functional database application.

Research Method

The iterative data modelling approach is also of value as a research method. The aforementioned distinction between the conceptual/logical level and the interface level works well when the data for a data model is complete and unambiguous and the process in which the data model plays a role is completely mapped out. Obviously, these variables rarely hold true for research projects in the humanities.

Oftentimes the data model does not correspond with data at hand. First, a data model may ask for data that is not there for the majority of data objects. Secondly, data may be too vague to fit typed fields defined in a data model. Thirdly, a data model may lead to a research process that is too time consuming due to its level of detail. In all these cases, revisions of the data model are needed in order to continue the research process.

Instead of smoothing out irregularities in the data by simplifying the data model, the model should be adjusted to reflect the existing complexities, vagueness, and uncertainties. As Rawson and Muñoz (2016) have stated, scholars should "see the messiness of data not as a block to scalability but as a vital feature of the world which our data represents and from which it emerges." We encourage scholars to include these data driven practices into their data model and have developed various strategies and features, such as 'reversed classification', to allow them to do this in nodegoat (van Bree and Kessels, 2014; van Bree and Kessels, 2017).

With the iterative methodology applied in nodegoat, we have facilitated research projects in the range of: disambiguation of Babylonian letters, questions of provenance and intertextuality in medieval manuscripts, creation of a multi-sourced 19th century context of conference attendance on social issues, mapping structures of violence in 1965 Indonesia, and documentation of an actor-network towards an encyclopedia of romantic nationalism.

An iterative data modelling approach allows scholars to enrich their data model during the research process. While a scholar may first want to use their own model, this can later be transformed or mapped to existing standards, like CIDOC-CRM, or semantic web standards. The data itself may be enriched by adding external identifiers such as VIAF identifiers or identifiers to other linked open data resources. This last point is important when data is published as an actionable dataset online (Berners-Lee, 2006).

Conclusion

In this paper we have set out to describe an iterative data modelling approach that helps scholars become confident in modelling their data and that functions as a research method for database development in the humanities. We have argued how a continuous shift between three levels of data modelling helps to conceive actionable datasets and establishes a framework for dealing with the complexities associated with humanities research.

Bibliography

Beretta, F. (2016). From Index Cards to a Digital Information System: Teaching Data Modeling to Master's Students in History. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 132-135.

Berners-Lee, T. (2006) Linked Data. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html.

Bree, P. van and Kessels, G. (2014) Reversed Classification [Blog Post]. Retrieved from https://nodegoat.net/blog.s/5/reversed-classification.

Bree, P. van and Kessels, G. (2015) “Mapping memory landscapes in nodegoat” in: *Social Informatics,* ed. L.M. Aiello and D. McFarland (Lecture Notes in Computer Science 8852), pp 274--278, New York : Springer International.

Bree, P. van and Kessels, G. (2017) Formulating Ambiguity in a Database [Blog Post]. Retrieved from https://nodegoat.net/blog.s/21/formulating-ambiguity-in-a-database.

Erickson, A. T. (2013). Historical Research and the Problem of Categories. In: Dougherty, J. and Nawrotzki, K. (eds), *Writing History in the Digital Age*. Ann Arbor: University of Michigan Press, pp. 133-145.

Flanders, J. and Jannidis, F. (2015) Data Modeling, in *A New Companion to Digital Humanities* (eds S. Schreibman, R. Siemens and J. Unsworth), John Wiley & Sons, Ltd, Chichester, UK. doi: 10.1002/9781118680605.ch16

Manovich, L. (1999). Database as a symbolic form. *Millennium Film Journal*, 34 (Fall)

Posner, M., (2015) Humanities Data: A Necessary Contradiction [Blog Post]. Retrieved from http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/.

Rawson, K. and Muñoz, T. (2016) Against Cleaning. Retrieved from: http://www.curatingmenus.org/articles/against-cleaning.

Defactoring ‘Pace of Change’: Exploring Code Review Methods for Textual Scholarship and Literary Studies

Joris J. Van Zundert
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands Royal Netherlands Academy of Arts and Sciences

Matt Burton
mcburton@pitt.edu
University of Pittsburgh, United States of America

Introduction

We start from the assertion that coding and code—as the source code of computer programs that is readable to humans and which drives the performative nature of software (Ford 2015, Hiller 2015)—can be inherent parts of scholarship or scholarship by and of themselves. That is: we assert that code can be scholarly, that coding can be scholarship, and that there is little difference between the authorship of code or text (Van Zundert 2016). The dichotomy that has been often sought between on the one hand a ‘pure’ intellectual realm associated with scholarly writing and academic print publication, and on the other hand the ‘material labour’ associated with for instance instrument making or programing, is artificial.

We argue the validity of this assertion along Burgess and Hamming (2011) and Clement (2016). These scholars refer to earlier work in which Bruno Latour (1993) casts the defining characteristic of modernity as a process of ‘purification’ which aims to contrast the human culture of modernity to nature. Burgess and Hamming observe a congruent process in academia: “Within the academy we see these processes of purification and mediation at work, producing and maintaining the distinction between intellectual labor and material labor, both of which are essential to multimedia production” (Burgess & Hamming 2011:¶11). This process serves to distinguish between scholarly and non-scholarly activities: “The distinction between intellectual and material labor is pervasive throughout scholarly criticism and evaluation of media forms. […] In addition, any discussion of scholarly activities in multimedia format are usually elided in favor of literary texts, which can be safely analyzed using traditional tools of critical analysis.” However, this distinction is based upon a technological fallacy already pointed out—as Burgess and Hamming note—by Richard Grusin in 1984. Grusin argued that Hypertext has not changed the nature of text essentially, as writing has always already been hypertextual through the use of indices, notes, annotations, and intertextual references. To assume that the technology of Hypertext has unvealed or revolutionary activated the associative nature of text, amounts to the fallacy of ascribing the associative agency of cognition to the technology, which however is of course a ‘mere’ expression of that agency.

Analogous to Burgess and Hamming, we argue that relegating the evaluation of scholarship to the reviewing of print publications is an equal fallacious ascribing of agency to the technology of written text. Such a narrow understanding of scholarship presupposes that something is scholarship because it is in writing, that writing makes it scholarship.

It is possible to evade all such possible technological fallacies by understanding scholarship as argument. We argue therefore that scholarship in essence is argument, and that technologies enable to shape and express that argument. This is not to say that technologies are mere inert and neutral epistemological tools, obviously different technologies shape and affect argument in different ways. Different technologies can therefore enrich scholarly argument. Scholarship is thus not bound to the use of text as an epistemological technology, but essentially is in the shaping of an argument. Text and writing may still be the most celebrated semiotic technologies to express an argument, but computer code understood as ‘just another’ literacy (cf. Knuth 1984, Kittler 1993, Vee 2013) can equally be the carrier of scholarly argument.

However, the acceptance of code as another form of scholarly argument presents problems to the current scholarly process of evaluation because of a lack of well developed methods for reviewing and critiquing scholarly code. Digital humanities as a site of production of non conventional research outputs—digital editions, web based publications, new analytical method, and computational tools for instance—has spurred the debate on evaluative practices in the humanities considerably, exactly because practitioners of digital scholarship acknowledge that much of the relevant scholarship is not expressed in the form of traditional scholarly output. Yet the focus of review generally remains on “the fiction of ‘final outputs’ in digital scholarship” (Nowviskie 2011), on old form peer review (Antonijevic 2016), and on approximating equivalencies of digital content and traditional print publication (Presner 2012). Discussions around the evaluation of digital scholarship have thus “tended to focus primarily on establishing digital work as equivalent to print publications to make it count instead of considering how digital scholarship might transform knowledge practices” (Purdy & Walker 2010:178, Anderson & McPherson, 2011). As a reaction digital scholars have stressed how peer review of digital scholarship should foremost consider how digital scholarship is different from conventional scholarship. They argue that review should be focused on the process of developing, building, and knowledge creation (Nowviskie 2011), on the contrast and overlap between the representationality of conventional scholarship and the strong performative aspects of digital scholarship (Burgess & Hamming 2011), and on the medium specificity of digital scholarship (Rockwell 2011).

The debate on peer review in digital scholarship however, has been geared much to high-level evaluation, concentrating for instance on the issue how digital scholarship could be reviewed in the context of tenure track evaluations. Very little has been proposed as to concrete techniques and methods for more practical level applied peer review of program code. Existing practical guidance pertains to digital objects such as digital editions (Sahle & Vogler 2014) or to code as cultural artefact (Marino 2006), but no substantial work has been put forward on how to peer review scholarly code. We are left with the rather general statement that "traditional humanities standards need to be part of the mix, [but] the domain is too different for them to be applied without considerable adaptation" (Smithies 2012), and the often echoed contention that digital artefacts should be evaluated as such and not as to how they might have been documented in conventional articles. The latter argument probably most succinctly put by Geoffrey Rockwell (2011): "While such narratives are useful to evaluators [...] they should never be a substitute for review of the work in the form it was produced in."

Yet, the problem is growing more urgent. Increasingly, code is created and used as a mechanism of analysis in textual scholarship and literary studies—cf. for instance Enderle 2016, Jockers 2013, Piper 2015, Rybicki et al. 2014, and Underwood 2014—which leads to the need to evaluate the technical, methodological and epistemological qualities of such code, as for instance the 'Syuzhet case' showed (Swafford 2016). The algorithms, code, and software that underpins the analyses in these examples of scholarship are not standardized 'off the shelf' software productions. These code bases are nothing like a software package or product such as AntConc that can be viewed as a generic and packaged distributable tool; a tool that might be subject to a scholarly type of tool criticism explaining and opening it for reuse by other scholars. Instead these codebases are bespoke code: they are one-off highly specific and complex analytical engines, tailored to solving one highly specific research question based on one specific set of data. Reuse, scalability, and ease-of-use are, justifiably (Baldrigde 2015), not specific aims of these code objects at all. Such might be the case with generic software, but these programs have been algorithmic instruments tailor made to serve the research case at hand. As such—and following what was argued above—we must regard these code bases as an inherent part of the scholarly argument they contribute to. And as such they deserve and require specific and rigorous peer review, like any argument in humanities research. How such peer review should be conducted is, however, a large unknown.

As a contribution to the challenges of code peer review we present an experimental technique we call defactoring. Drawing on Braithwaite (2013), we have re-configured the program code that underpins a recent article by Ted Underwood and Jordan Sellers (Underwood & Sellers 2016) into a computational narrative—echoing Knuth's literate programming (1984)—to be critically analyzed and annotated. This method is intimately intertwined with the Jupyter Notebook platform, which allows for the composition of scholarly and scientific inscriptions that are simultaneously human and machine readable. The Notebook is both a document format and a platform for mixing code and prose into executable objects. We have extracted Underwood and Seller's code and defactored it into a Jupyter Notebook, available on Github. This means we have recombined code from disparate files, linearized the execution path, demodularized function calls, and annotated code blocks with our own expository comments. As an annotated Notebook we can now engage Underwood and Seller's code directly as a scholarly inscription and more deeply interrogate the role of data, algorithms, and code in the production of knowledge.

In the case of scholarship that uses computation, large parts of the intellectual importance are embodied in the code rather than living exclusively in the print publication. As our case study also shows, usually the method description in the print publication presents the intellectual contribution of the code development in a very reduced, and rather imprecise high-level fashion. The code itself is a more precise inscription of the analysis the researchers conducted. The methodological approach we present is a way to engage the code, and thus allows a peer reviewer to understand and interpret the intellectual narrative of the code. This results in a fuller grasp and understanding of the methods applied, and thus to a more comprehensive review of the intellectual effort associated with the publication.

After demonstrating the work in the notebook, we will conclude our paper with a critical reflection of the reviewing work that was undertaken with it. We identify the applicability, feasibility, benefits, and drawbacks of this specific approach. We also outline some possible future directions of research that could further contribute to exploring review methods for code scholarship.

Bibliography

Anderson, S., McPherson, T., (2011). Engaging Digital Scholarship: Thoughts on Evaluating Multimedia Scholarship. Profession 136–151.

Antonijevic, S., (201)5. Amongst Digital Humanists: An ethnographic study of digital knowledge production, Palgrave Macmillan.

Baldridge, J., (2015). It's okay for academic software to suck. Java Code Geeks. Available at: https://www.javacodegeeks.com/2015/05/its-okay-for-academic-software-to-suck.html [Accessed April 25, 2016].

Braithwaite, R., (2013). Defactoring. Reginald Braithwaite: via raganwald.com. Available at: http://raganwald.com/2013/10/08/defactoring.html [Accessed March 15, 2016].

Burgess, H.J. & Hamming, J., (2011). New Media in Academy: Labor and the Production of Knowledge in Scholarly Multimedia. DHQ: Digital Humanities Quarterly, 5(3). Available at: http://digitalhumanities.org/dhq/vol/5/3/000102/000102.html [Accessed September 2, 2016].

Clement, T.E., (2016). Where Is Methodology in Digital Humanities? In Debates in the Digital Humanities 2016. University of Minnesota Press, pp. 153–175. Available at: http://dhdebates.gc.cuny.edu/debates/text/65.

Enderle, J.S., (2016). A Plot of Brownian Noise. Jonathan Scott Enderle. Available at: https://github.com/senderle/svd-noise/blob/master/Noise.ipynb [Accessed September 24, 2016].

Ford, P., (2015). What is Code? Businessweek. Available at: http://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/.

Grusin, R., (1994). What is an Electronic Author? Theory and the Technological Fallacy. Configurations, 2(3), pp.469–483.

Hiller, M., (2015). Signs o' the Times: The Software of Philology and a Philology of Software. Digital Culture and Society, 1(1), pp.152–163.

Jockers, M.L., (2013). Macroanalysis: Digital Methods and Literary History, Urabana, Chicago, Springfield: UI Press.

Kittler, F., (1993). Es gibt keine Software. In Draculas Vermächtmis. Leipzig: Reclam Verlag, pp. 225–242.

Knuth, D.E., (1984). Literate Programming. The Computer Journal, 27(1), pp.97–111.

Latour, B., (1993). We Have Never Been Modern, Cambridge, Massachusetts: Harvard University Press.

Marino, M.C., (2006). Critical Code Studies. Electronic Book Review. Available at: http://www.electronicbookreview.com/thread/electropoetics/codology [Accessed January 16, 2015].

Nowviskie, B., (2011). Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. Profession, pp.169–181.

Piper, A., (2015). Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel. New Literary History, 46(1), pp.63–98.

Presner, T., (2012). How to Evaluate Digital Scholarship. Journal of Digital Humanities, 1(4). Available at: http://journalofdigitalhumanities.org/1-4/how-to-evaluate-digital-scholarship-by-todd-presner/.

Purdy, J.P. & Walker, J.R.,(2010). Valuing Digital Scholarship: Exploring the Changing Realities of Intellectual Work. Profession, pp.177–195.

Rockwell, G.,(2011). On the Evaluation of Digital Media as Scholarship. Profession, pp.152–168.

Rybicki, J., Hoover, D. & Kestemont, M., (2014). Collaborative authorship: Conrad, Ford and Rolling Delta. Literary and Linguistic Computing, 29(3), pp.422–431.

Sahle, P. & Vogeler, G., (2014). Criteria for Reviewing Scholarly Digital Editions (version 1.1). Institut für Dokumentologie und Editorik. Available at: http://www.i-d-e.de/publikationen/weitereschriften/criteria-version-1-1/ [Accessed October 13, 2016].

Smithies, J., (2012). Evaluating Scholarly Digital Outputs: The Six Layers Approach. Journal of Digital Humanities, 1(4). Available at: http://journalofdigitalhumanities.org/1-4/evaluating-scholarly-digital-outputs-by-james-smithies/ [Accessed September 2, 2016].

Swafford, J. (2016) 'Messy Data and Faulty Tools', in Gold, M. K. and Klein, L. F. (eds) Debates in the Digital Humanities. Minneapolis: University of Minnesota Press, p. 600. Available at: http://dhdebates.gc.cuny.edu/debates/text/100.

Underwood, T.,(2014). Understanding Genre in a Collection of a Million Volumes, Interim Report. Figshare. Available at: https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251 [Accessed March 15, 2016].

Underwood, T. & Sellers, J.,(2016). The Longue Durée of Literary Prestige. Modern Language Quarterly, 77(3), pp.321–344.

Vee, A., (2013). *Understanding Computer Programming as a Literacy*. LiCS, 1(2), pp.42–64.

Zundert, J.J. van, (2016.) *Author, Editor, Engineer — Code & the Rewriting of Authorship in Scholarly Editing.* Interdisciplinary Science Reviews, 40(4), pp.349–375.

Less intent, more impact: Transforming public DH projects toward access, care, and inclusion

Amanda Visconti
amandavisconti@gmail.com
University of Virginia, United States of America

Good intent carries no guarantee of a positive impact on the world. For digital humanities designers and makers building projects that face or involve public audiences, we have an extra scholarly and human responsibility for all the repercussions of our work. Focusing not on what we mean to create, but on how our work could both positively and negatively affect others, enables a more caring, accessible, and inclusive DH.

This paper uses my design, coding, and user testing of the participatory digital edition of James Joyce's challenging novel Ulysses, *Infinite Ulysses* (InfiniteUlysses.com), to demonstrate a public DH project designed with care for its impact on both its audience and the DH community. More importantly, I'll explore how ethical design considerations such as increasing accessibility and inclusion can be added or increased in existing projects that may not have explicitly built these values in from the start.

Infinite Ulysses is a public digital humanities project that has drawn over 25,000 unique site visitors, as well as a smaller body of 775 readers who created user accounts to interact more closely with the digital edition through annotation and other social features. *Infinite Ulysses* is also the bulk of a unique, successfully defended, no-chapters literature dissertation, and as such those interested in the future of humanities graduate education may enjoy this paper.

Building the *Infinite Ulysses* digital edition helped me separate the scholarly values of textual scholarship from the common embodiment of these values (i.e. the scholarly digital edition or SDE). Through this clarification, I imagined new types of digital edition that, while different from SDEs, hold true to the same values performed by SDEs. Through building and testing the public use of one of these new models for reifying textual scholarship values, I experimented with designing an edition that is not just publicly accessible, but also invites and assists public participation in the scholarly love for a text's materiality, history, and meaning. On the spectrum of crowdsourcing engagement, I located multiple paths for meaningful public DH activity that fall between the endpoints of full critical rhetoric and "adding a tag". To enable this public engagement, I explored ways of designing participatory digital editions to adeptly handle an influx of public readers and their annotations, and tested suggestions on what scholars can learn about digital editions and their texts from the accompanying influx of website use data.

The Femtechnet scholars' recentering of DH on ethical questions around technology design led me to connect the fields of textual scholarship and human-computer interaction, porting concepts of humane and ethical digital design to the domain of literary editions. Michael Muller's formulation of participatory design as a "third space" connecting two audiences helped me support both scholarly and public edition readers in "challenging assumptions, learning reciprocally, and creating new ideas, which emerge through negotiation and co-creation of identities, working languages, understandings, and relationships, and polyvocal (many-voiced) dialogues across and through differences". Katie Shilton's exploration of how the design of a technology shapes the social values and ethics of its users helped me plan toward a community of annotators that would care for rather than compete against one another.

My post-dissertation work on *Infinite Ulysses* pulls in thinking from libraries and information science. Librarian Chris Bourg's argues that neither technology nor knowledge infrastructures such as libraries can be neutral, and that therefore we must leverage both toward, rather than away from, social justice. Archivist Jarrett M. Drake calls us to move beyond current institutional archives to build knowledge structures that let us "unlock our futures as humans, as community members, as archivists, and as memory workers", and suggests design that encourages perceptions of community belonging as a way forward. Digital Library Federation Director Bethany Nowviskie asks, "Can we position our digital collections and digital scholarly projects more plainly not as statements about what was and is, but as resources for the building of different, better worlds?" The knowledge infrastructure of libraries and archives ports well to that of digital editions, as well as to DH more broadly. Each of these thinkers push us to create work that would take us from what DH is to what DH can be; in this presentation, their arguments act as direct prompts for exploring small ways of designing toward a better DH through the specific use case of *Infinite Ulysses*.

This paper quickly but clearly paints a cohesive picture of the *Infinite Ulysses* project, instead spending the bulk of its time critiquing the impact of the project on its users and on the DH community. I will survey the critical DH cultural scholarship that grounded my assessment of the ethical impact of my work (e.g. Roopika Risam, Moya Bailey, Amy Earhart), as well as scholarship in related areas such as libraries feminist interface design (e.g. Bess Sadler, Catherine D'Ignazio and Lauren F. Klein, Shaowen Bardzell and Eli Blevis).

In "Do Artifacts Have Ethics?", Michael Sacasas proposes a series of questions to be asked by anyone making something of what they're making—whether that's building an app, designing a class, or in this case designing and coding a digital humanities project. I use Sacasas' provocations to examine the successes, failures, and difficult decisions of *Infinite Ulysses*' design, walking the audience through concrete examples of what I did, what I should have done, or what I could do to make the project have an even better impact on the world.

- *An example of a positive impact on Infinite Ulysses' users:* I provided my annotators with the ability to export their textual annotations from my website, giving them ownership and control over the results of their labor, and not forcing them to risk losing their work if my site crashed or closed.
- *An example of a negative impact on Infinite Ulysses' user*s: I tried to populate the digital margins of Ulysses with the characters in the book, by letting new users choose among illustrations of the novel's characters for the avatar that accompanies their textual annotations. The makeup of Joyce's text means that those avatars were largely white and male, with three white female options, a cat, and no other identities. For a text that already struggles with instances of sexism, racism, and transphobia, making new readers feel "I'm not represented here, I'm not supposed to be here" was an additional negative impact.
- *An example of a positive impact of Infinite Ulysses on the DH community:* I demonstrated that a public DH project could support meaningful public participation that didn't necessitate bending public questions, interpretations, and other comments on Ulysses to fit the mold of scholarly rhetoric. By both successfully designing for public participation (my edition was cited in The New York Times) and performing innovative humanities scholarship through that design (the

project received my university's 2016 award for arts and humanities distinguished dissertation), I added to the digital humanities' ever-growing examples of how building can itself be critical research.

- *An example of a negative impact of Infinite Ulysses on the DH community:* I deepened digital editions' duplication of the problems with the print literary canon, by building a project around another canon text and author. When only 11 of 86 projects funded by the NEH Preservation & Access Office 2006-2011 had a topic other than a white male writer, I struggle with whether the positive public impacts of my work are worth deepening our problems with authorial representation.

I act as my scholarship's own harshest critic not to paralyze other DH builders from ever making anything, but to make visible specific examples of how in small ways, by adding the skill of care to the skill of critique, we makers can build a better DH and a better world.

Bibliography

Arbuckle, A. (2014). "Considering The Waste Land for iPad and Weird Fiction as models for the public digital edition". Digital Studies. 2014.

Bailey, M. (2015). "#transform(ing)DH Writing and Research: An Autoethnography of Digital Humanities and Feminist Ethics". *Digital Humanities Quarterly* 9(2). digitalhumanities.org/dhq/vol/9/2/000209/000209.html

Bardzell, S. and Blevis, E. (2010) "The lens of feminist HCI in the context of sustainable interaction design". interactions 17(2). March-April 2010: 57-59.

Bourg, C. (2016) "Libraries, technology, and social justice". Text of Access 2016 talk published as blog post on Feral Librarian blog. October 7, 2016. chrisbourg.wordpress.com/2016/10/07/libraries-technology-and-social-justice

D'Ignazio, C. and Klein, L. F. (2016) "Feminist Data Visualization". 2016 IEEE VIS conference paper. kanarinka.com/wp-content/uploads/2015/07/IEEE_Feminist_Data_Visualization.pdf

Drake, J. M. (2016) "Liberatory Archives: Towards Belonging and Believing (Part 1)". Medium blog post: "On Archivy" collection. October 22, 2016. medium.com/on-archivy/liberatory-archives-towards-belonging-and-believing-part-1-d26aaeb0edd1#.fbgxfse1v

Drucker, J. (2010). "Graphesis: Visual Knowledge Production and Representation". Poetess Archive Journal. 2(1). 2010. journals.tdl.org/paj/index.php/paj/article/download/4/50

Drucker, J. (2011). "Humanities Approaches to Graphical Display". Digital Humanities Quarterly. 5(1). 2011. digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Earhart, A. (2012). "Can Information Be Unfettered? Race and the New Digital Humanities Canon". From Gold, Matthew K., ed. Debates in the Digital Humanities. 2012 print edition as accessed online at dhdebates.gc.cuny.edu/debates/text/16.

Galey, A. and Ruecker, S. (2010) "How a Prototype Argues." *Literary and Linguistic Computing*. 25(4). 405-424.

Groden, M. (2001) "Introduction to 'James Joyce's Ulysses in Hypermedia'". *Journal of Modern Literature*. 24(3/4). 359-362.

Groden, M. (2003-4_ "Problems of Annotation in a Digital Ulysses" . *Hypermedia Joyce Studies.* 4(2). 2003-4. www.oocities.org/hypermedia_joyce/groden.html

Kraut, R. E. and Resnick, P. (2011) *Building successful online communities: evidence-based social design*. Cambridge, MA: MIT Press..

Losh, E., Wernimont, J., Wexler, L., and Wu, H.-A. (2016)"Putting the Human Back into the Digital Humanities: Feminism, Generosity, and Mess". *In Debates in the Digital Humanities*, Lauren F. Klein and Matthew K. Gold, eds. dhdebates.gc.cuny.edu/debates/text/61

Marino, M. C. (2007) "Ulysses on Web 2.0: Towards a Hypermedia Parallax Engine". James Joyce Quarterly. 44(3). Spring 2007. 475-499.

Matienzo, M. A. (2015) "To Hell With Good Intentions: Linked Data, Community and the Power to Name". Blog post of November 15, 2015 keynote for the 2015 LITA Forum. Matienzo.org/2016/to-hell-with-good-intentions

Muller, M. J. (2012) "Participatory Design: The Third Space in HCI". In *The Human-Computer Interaction Handbook*. 1051-1068.

Norris, M. (2011). *Virgin and Veteran Readings of Ulysses.* Palgrave Macmillan.

Nowviskie, B. (n.d.) "Interfacing the Edition". Text of conference talk.www2.iath.virginia.edu/bpn2f/1866/interface.

Nowviskie, B. (2016) "alternate futures/usable pasts". Text of talk given at Marquette University Library in September 2016 published as blog post. October 24, 2016. nowviskie.org/2016/alternate-futures-usable-pasts/

Risam, R. (2015) "Beyond the Margins: Intersectionality and the Digital Humanities". *Digital Humanities Quarterly* 9(2). 2015. digitalhumanities.org:8081/dhq/vol/9/2/000208/000208.html

Sacasas, L.M. (2014) "Do Artifacts Have Ethics?" *Technology, Culture, and Ethics* blog, November 29, 2014. thefrailestthing.com/2014/11/29/do-artifacts-have-ethics

Sadler, B. and Bourg, C. (2015). "Feminism and the Future of Library Discovery". Code 4 Lib 28. April 15, 2015. journal.code4lib.org/articles/10425

Shilton, K. (2013). "Values Levers: Building Ethics Into Design". Science Technology Human Values 38(3), May 2013, 374-397.

Utell, J. M. (2008) The Archivist, the Archaeologist, and the Amateur: Reading Joyce at the Rosenbach". Journal of Modern Literature. 31. 2008. 53-65.

Warwick, C., Terras, M., Huntington, P., and Pappa, N. (2008). "If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data". Literary and Linguistic Computing. 23(1). 2008. 85-102.

Dead and Beautiful: The Analysis of Colors by Means of Contrasts in Neo-Zombie Movies

Niels-Oliver Walkowski
walkowski@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften Germany

Johannes Pause
johannes.pause@tu-dresden.de
Technische Universität Dresden, Germany

The analysis of color in movies is a topic of increasing interest in the still-young research field known as digital film analysis or *cinemetrics*. The thesis of Brodbeck (2011) is one of the early examples which have been discussed broadly. Brodbeck created a donut plot for each movie in which dominant colors of scenes are represented by colored slices within the plot. Similar approaches were taken by (Baker 2015) and (Burghardt, Kao, and Wolff 2016). Both approaches identify dominant colors of all frames in a movie and represent these colors as lines in a barcode-like visualization. In the first and the third project dominant colors are created by means of a clustering-algorithm, more precisely *k-means*. This algorithm is one of the most common approaches in *color quantization* and is also delivered for such tasks in well-known computer vision libraries like OpenCV.

However, there are several issues with using k-means for color quantization in digital film analysis which are rarely discussed within the community. One of the more obvious problems is that k-means (and other non-hierarchical cluster algorithms) require pre-definition of the number of clusters to be found. Thus, an automated analysis of around 180,000 frames in a movie does not respect the fact that there are frames which are more complex in color than others. A more sophisticated procedure exists in which the k-means algorithm is used in a loop with different numbers of cluster. In this approach the best fitting number of clusters is defined and evaluated by the *silhouette coefficient* which is applied to the result of each loop. However, Figure 1 demonstrates that the best-fit in terms of clustering is definitely not the best-fit for the interpretation of colors in film analysis.

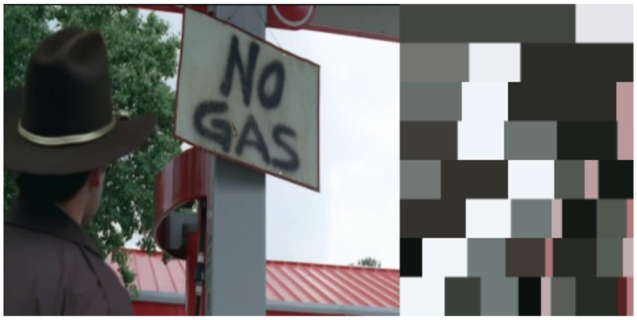


Figure 1: Cluster analysis of dominant colors in a frame image of *The Walking Dead.* According to the silhouette coefficient the number of two clusters produces the best clustering result.

Colors which seem to be important even before any interpretation has taken place disappear in the clustering process because they are 'eaten' by other clusters. Figure 2 offers a very impressive example for this phenomenon. The red girl from *Schindler's List* – the most exciting thing that happens in the movie in terms of color– is not represented within the clusters transparently. The reason is that k-means has a tendency towards equally sized clusters and that the color difference is not big enough to survive this tendency.

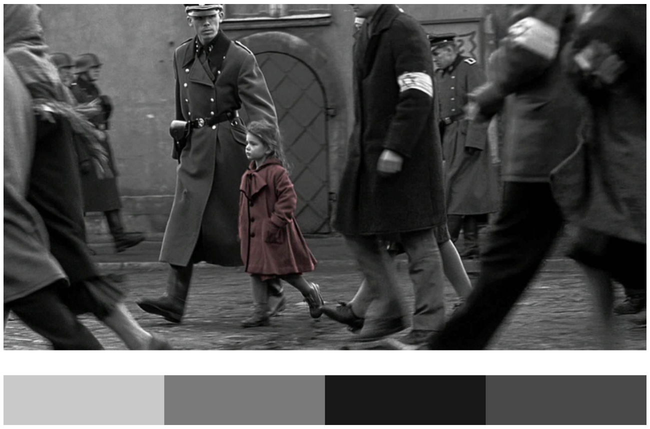


Figure 2: Cluster analysis of color in a scene from 'Schindler's List'

Another problem is that k-means produces slightly different results each time it runs. Hence, sometimes a color which would be perceived as different by humans is present in the result and sometimes it is not. The results differ most in between three and five clusters as predefined cluster values. However, this is exactly the span which is most often used for color quantization in digital film analysis. Finally, the k-means algorithm produces different results in correspondence with the color space in which a frame is represented. For instance, the HSV color space produces better results than the RGB color space.

All of these problems call for a theoretical evaluation of what dominance means for dominant colors in digital film analysis. Unfortunately, the previously mentioned projects have not presented such evaluations. Furthermore, these projects also did not systematically interpret the results they have produced. For this reason, it is not transparent which semantics can be found in the patterns.

All this being said, a different approach for the computational analysis of colors in movies seems necessary. This approach needs to introduce a starting point which is technologically less arguable and which offers concepts for the interpretation of results. The approach that will be presented builds upon the theory of *seven color-contrasts* developed by Johannes Itten (Itten 1961, 36–109). Itten claims that the effect of colors is not absolute but depends on the surrounding color environment. Colors interact with each other and there are seven contrasts in which this interaction can be analyzed. These contrasts are:

- hue
- saturation
- light-dark
- cold-warm
- complementary
- simultaneous
- extension

Each contrast has certain capabilities to structure and create effects in the narrative and aesthetic design of a movie. They can be used: to guide the attention of viewers, to create spatial ambiance, to create orientation or confusion, to support the symbolic layer or to create associations as well as emotions. Although these effects are not generalizable the approach of color-contrasts has more to provide to interpret color in movies than the analysis of single colors. This holds true especially where the dynamics of multiple contrasts are related with each other and begin to form a language of color. For instance, a movie might have a stable opposition between cold and warm colors but more progressive color dynamics (Wulff 1988) between light and dark colors. The results of such an analysis can be related to narrative aspects, characters, leitmotifs or the *Mise en Scène* in a movie or a corpus of movies.

The technological implementation of this approach depends on the type of contrast that will be analyzed. Data for the first three contrasts can be obtained by converting movie frames to the HSV or HSB color space. The conversion between the YUV color space (in which many movie files are represented) into HSV is lossless in most cases (Ford and Roberts 1998). Each channel in HSV represents one such contrast. Usage of the CIE L*a*b* color space can be considered to comply with certain issues of color awareness that are not tackled by HSV and others. Other contrasts require further processing. For instance, to obtain cold and warm color values, each value in the hue channel can be associated with a value that represents how warm or cold it is. In general, values of red and yellow are conceived as warm. However, this effect is very much influenced by cultural and psychological aspects (Küppers 2000). Therefore, the association of color values with values of warmth is a task that requires decision-making. In contrast, the presence of complementary and simultaneous colors is a mathematical relationship and can be computed within a certain color space consistently.

There are also several ways in which contrast data can be analyzed. Figures 3 and 4 visualize two of such strategies. In both cases a histogram of one contrast was computed for every second of the movie. In Figure 3 the leftmost and rightmost peaks in each histogram was calculated. These peaks constitute the min- and max-bound of the contrast span in each frame. The x-axis represents the time-axis of the movie while the y-axis shows the value of each min- and max-bound.

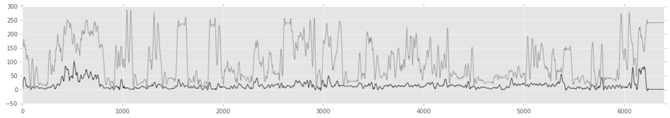


Figure 3: Slightly interpolated boundaries of light-dark contrast in '28 Days Later'

In Figure 4 the contrast of a spectrum was reduced to 16 bins and each contrast value that was produced by a critical number of pixels was plotted as a point. The size of the point represents the number of pixels in a frame for each bin.

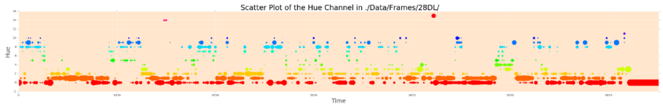


Figure 4: The appearance of hue values above a certain threshold in '28 Days Later'

The movie which underlies both visualizations is *28 Days Later*. Figure 3 represents the light-dark-contrast. It shows two sequences in which light and dark colors go up simultaneously for a certain amount of time. One sequence is between the 200th and the 800th frame the other between the 6000th and the beginning of the credits. In the first sequence the protagonist Jim awakes from coma and realizes that the world fell apart. In the second sequence the main group of persons reach the final save place. Thus, a similar pattern in one contrast frames the actual storyline. However, the color-itself-contrast represented in Figure 4 is extremely different in these scenes. The first scene has a narrow spectrum while the spectrum of the second scene is broad. The narrow spectrum provokes disorientation because it tones down differences. The color segment is literally dazzling. In contrast, the spectrum and coverage of colors at the end mediate clarity, stability and order.

This presentation will outline the problems of k-means for color quantification in digital film analysis. It will describe the theory of seven color-contrasts and give examples how such theory can be adopted computationally. Each step will be illustrated by an analysis of a corpus of three neo-zombie-movies, more precisely *28 Days Later*, '[REC]' *and World War Z*.

Bibliography

Baker, D. (2015). “Spectrum.” *Dillon Baker. http://dillon-baker.com/spectrum/*

Brodbeck, F. (2011). “CINEMETRICS Film Data Visualization.” *http://cinemetrics.fredericbrodbeck.de/*

Burghardt, M., Kao, M., and Wolff, C.. (2016). “Beyond Shot Lengths and Color Information as Additional Parameters for Quantitative Movie Analysis.” In *Conference Abstracts*, 753–55. Kraków: Jagiellonian University & Pedagogical University

Ford, A., and Roberts, A. (1998). “Colour Space Conversions.”

Itten, J. (1961). *Kunst Der Farbe*. Ravensburg: Otto Maier Verlag.

Küppers, H. (2000). *Harmonielehre Der Farben. Theoretische Grundlagen Der Farbgestaltung*. 2nd ed. Köln: DuMont

Wulff, H. J. (1988). “Die Signifikativen Funktionen Der Farben Im Film.” *Kodikas/Code* 11 (3-4): 363–7

Livingstone Online: Access Beyond Openness

Megan Ward
megan.ward@oregonstate.edu
Oregon State University, United States of America

Adrian S. Wisnicki
awisnicki@yahoo.com
University of Nebraska - Lincoln,
United States of America

The study of nineteenth-century Africa is troubled by issues of access on two fronts. First, explorers’ unedited field notes – the closest thing we have to a “‘raw’ record” – are rarely available, and, if they are, they are often crumbling, illegible, or located in far-flung archives (Bridges 1998). Second, even when such sources are available in later published forms, they present ethical and ideological problems. Written largely by European explorers and heavily edited, the published texts often exclude the voices of the very populations to which they attempt to provide access. *Livingstone Online*, a digital project dedicated to the written and visual legacy of nineteenth-century explorer David Livingstone (1813-73), works to counter these issues through its site design, transcription processes, and use of spectral imaging technology.

In order to do so, however, we have had to reconsider our understanding of access, both technologically and ideologically. As a publicly-funded project, we adhere to a high standard of transparency. Yet, as an archive of a contentious figure of imperial exploration, we are also responsive to the recent critiques of access - both of open access as privileging imperial knowledge expansion (Christen 2012; Risam 2017) and of the digital humanities as excluding consideration of race (Gallon 2016). To navigate this conflict, we strive to provide access that is not simply based on openness.

Instead, our project offers an understanding of access that moves in two directions temporally: striving to repair the past by being ethical in our digital treatment and remediation of historical materials, while also acting in a future-oriented fashion in developing and implementing our transparency policies, data standards, and code of collaboration in order to engage a variety of audiences, including those often excluded from DH practice. In this way, our project attempts to create a digital platform for culturally sensitive materials, while our documentation procedures seek to reveal every step of our decision-making process to critical review.

Reparative

Livingstone Online, now in its twelfth year (2004 present), is a digital museum and library that draws on recent scholarship and international collaboration to restore one of the British Empire's most iconic figures to his global contexts. Our digital collection of high-resolution manuscript images and critically-edited transcriptions – 11,000 images and 700 transcriptions by 2017 – is among the largest on the internet related to any single historical British visitor to Africa. Our site publishes important research on Livingstone's legacy and explores the many ways his ideas have circulated over time. Uniquely, we also takes its visitors far behind the scenes of our work – documenting step-by-step the international collaboration among archives, scholars, scientists, librarians, computer programmers, and other specialists that has made our project possible.

Our use of spectral imaging to uncover the material history of Livingstone’s manuscripts gives us important insights into the conditions under which Livingstone and other imperial explorers wrote – from unacknowledged contributors to the many environments through which the manuscripts circulated. In foregrounding these dimensions, we are also creating a new approach to using spectral imaging in cultural heritage projects because spectral imaging has primarily been used to unearth layers of text, rather than to examine the broader circumstances of imperial record-making and the preservation of imperial records over time. This new use of spectral imaging also constitutes an ethics of access, which is framed by critical essays that explore Livingstone’s uncredited information sources. Livingstone Online here puts forward an idea of access as uncovering the hidden hands and voices of the past.

For instance, in our study of Livingstone’s 1871 Field Diary, we have collaborated with spectral imaging scientists to develop pseudocolor (false color) images to differentiate passages that Livingstone originally wrote from those he added latter and those added by other hands. Likewise, the development of animated spectral images has enabled a chronological reconstruction of events in the life of the diary. By contrast, recourse to images made by principal component analysis (PCA) has uncovered stains on pages otherwise invisible to the naked eye and has introduced us to dimensions of manuscript history otherwise not even suspected to exist.

In addition, we frame these spectral images with paratextual tools that value equally the different kinds of information that Livingstone records. For example, few or no other record remains of many of the villages or the African and Arab individuals Livingstone mentions. As a result, our integrated glossary offers unique, otherwise unavailable geographical information that circulated during Livingstone's time in Central Africa and enumerates the names of people that might otherwise be lost to history. The glossary and other critical materials also provide insight into the complex social dynamics that operated in areas where Livingstone traveled. Overall, Livingstone Online offers a version of access in which the freely available manuscript pages are only a starting point; spectral imaging technology combined with critical building helps construct a reparative ethos of access.

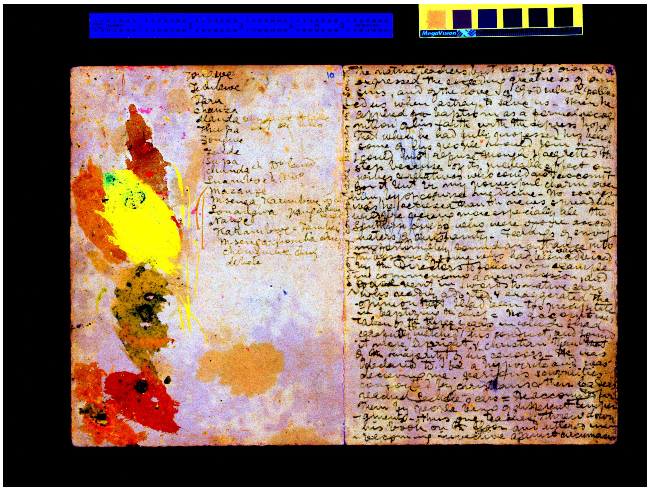
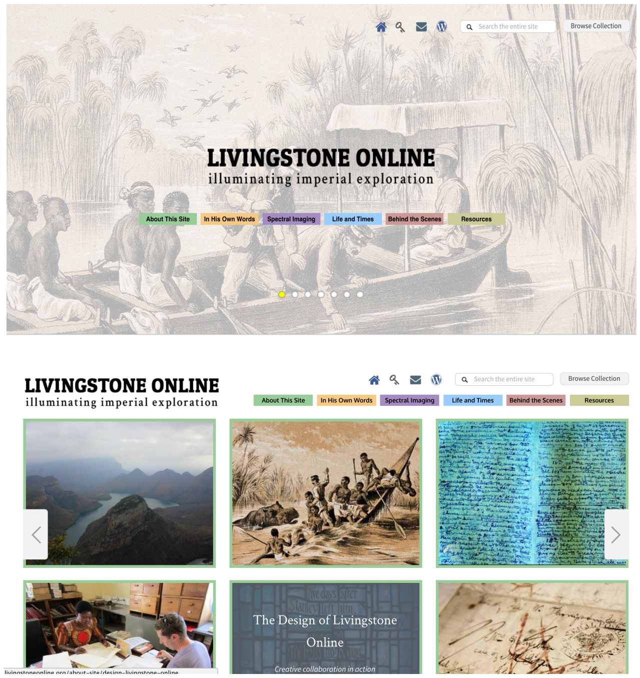


Figure 1. A processed spectral image of a two-page spread from the 10 March 1870 'Retrospect' (Livingstone 1870a:[3]). The kaleidoscopic colors foreground and differentiate the wide range of substances that have left traces on the manuscript's pages over its 145-year history. Copyright National Library of Scotland. Creative Commons Attribution-NonCommercial 3.0 Unported

Future-Oriented

Alongside such reparative work, we also strive to design a project that looks toward the future. Livingstone Online makes our project documents available to an almost unprecedented degree in order to make our publicly-funded research fully accountable, to illuminate our work practices, and to support future digital projects. Our extensive downloadable primary materials (including 12,000 images, 3000 metadata records, and hundreds of transcriptions) are supplemented by freely available project materials, images, and working documents – the things often hidden behind the public face of digital projects. We have curated access to over 600 project documents, including planning documents, spectral image processing information, and essay notes, in order to illuminate the long-term history of our project work. Likewise, access to our grant narratives and working documents de-mystifies funding processes and international, interdisciplinary collaboration in order to support the work of other scholars, especially junior or independent scholars and those new to DH.

In addition, our site is technically accessible in a range of ways. We have built the site with sustainable, community-supported, open-source technologies such as Drupal for our front end and Fedora for our back end, which means that others can access our underlying code (which is fully available from Github) to reuse and modify it for their own projects. To promote use of our site more broadly, we've also worked to make the site inviting to scholars and general users alike, using intuitive, visually-driven site design. Through our site design, open-source code, and transparent documentation, we hope to foster user-led interpretation over passive reception of authorized knowledge.



Figures 2,3. Livingstone Online's six section pages, two of which are pictured here, each rely on a diverse range of historical illustrations and contemporary images to complicate the notion of a definitive Livingstone.

As part of this effort, our site is also fully mobile accessible, including for complex functions such as the review and study of archival manuscripts and transcriptions. This opens up Livingstone's documents and our critical materials to the parts of the world where he worked and travelled; for many people on the African continent, for instance, mobile technology is the main access point to the internet. Likewise, just under a quarter of our collaborating archives are in Africa, and we are actively working on developing additional relationships with African-based archives, in the interests of not only bringing new Livingstone materials into our site, but also to encourage collaboration with African-based scholars and general audiences.

In these ways, we hope to initiate a conversation about the biases and assumptions inherent in the ways that technological advances shape our preservation of the past. We also hope to develop a nuanced practice of access that is embedded in our site design, spectral imaging processing, and transparent documentation, as well as made explicit in our critical materials. Thinking of access beyond openness means creating more historically-minded digital collections that also look to future knowledge creation by an array of populations, not all of them academic or based in the west.

Bibliography

Bridges, R. (1998). "Explorers' Texts and the Problem of Reactions by Non-Literate Peoples: Some Nineteenth-Century East African Examples." *Studies in Travel Writing* 2: 65-84.

Christen, K. (2012). "Does Information Really Want to be Free? Indigenous Knowledge Systems and the Question of Openness." *International Journal of Communication*. Web.

Gallon, K. (2016). "Making a Case for the Black Digital Humanities." In *Debates in the Digital Humanities*, eds. Lauren F. Klein and Matthew K. Gold. Univeristy of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/55

Risam, R. (2017, *forthcoming*). "Decolonising Digital Humanities in Theory and Practice." Routledge *Companion to Digital Humanities and Media Studies*. Ed. Jentary Sayers. London: Routledge.

Multiplying Access: the Marianne Moore Digital Archive's Tools and Methods for Collaboration

Nikolaus Wasmoen
nlwasmoe@buffalo.edu
University at Buffalo (State University of New York)
United States of America

Introduction

The Marianne Moore Digital Archive (MMDA) has begun to publish digital editions of the 122 manuscript notebooks of Modernist poet Marianne Moore. The notebooks contain a plethora of materials, include reading notes, recorded conversations Moore participated in or overheard, drafts of poems, travel descriptions, financial records, notes on concerts, lectures, classes, and sermons Moore attended, and other miscellanea documenting an active and increasingly prominent life as part of New York's literary and cultural scene from the 1910s to the 1960s. These unique resources pose significant resistance to accessing the trove of literary and cultural data they contain: they have never been published in any form, can be visited by appointment at a library in Philadelphia for only limited hours during the week, most of the notebooks are written in a cramped, uneven hand with frequent abbreviation and shorthand, Moore repurposed calendars and other cheap or free items that include pre-printed material interspersed with her writing, many notebooks are now too fragile for normal handling, and previous attempts at conservation have led to the disordering of many pages. The unfamiliar and sometimes obscure historical and cultural references that are needed to trace the broad scope of Moore's readings and activities present an additional layer of difficulty for most of the potential audience that could take advantage of the notebooks for literary, historical, or cultural research. In response to these challenges, the MMDA has created a team of Moore scholars and scholarly digital editors at three universities and partnered with the Center for Unified Biometrics and Sensors (CUBS), an automated handwriting recognition and machine learning lab at the University at Buffalo.

This paper reflects on our efforts to increase access with respect to the notebooks in several senses: design choices with respect to the digital editorial apparatus needed to make the notebooks usable to non-specialists; the implementation of a HubZero 2.0 collaborative hub platform, designed for use in the sciences and engineering, for a humanities project; and the development of a customized, integrated editor/viewer based on existing digital manuscript editing tools that can enable non-technical editors to participate more directly in the digital workflow. The hub platform and workflow solution are primarily designed to support the specific research needs of the MMDA and its users, however the project code and our methodologies can be readily applied to other digital editing projects and digital humanities collaborations.

Abstract

Marianne Moore (1887-1972) was among the foremost modernist poets of the early twentieth- century. Her work contributed to the revolution in poetic form and conceptions of poetry occurring during the 1910s and 1920s and she remained a poet of profound reflection and innovative design throughout her lifetime. In particular, Moore was among the first to conceive the poem as constructed through both language and visual design on the page and she was the first Anglo-American poet to divorce the poetic line from syntax. While her peers such as T. S. Eliot, Ezra Pound, William Carlos Williams, Wallace Stevens, and H.D. often excoriated each others' work, all were profound admirers of Moore's thought and poetry (see, for example, Leavell 2014, and Miller, 1995 and 2005). Moore was also significant to modernism in her decades-long reviewing of work by her contemporaries, her publication of essays on aesthetic and cultural topics, and in her editing of one of the premier periodicals of modernist literature and art, *The Dial*, from 1925-1929. Moore received the Pulitzer Prize, the Bollingen Prize, the National Book Award, and the Gold Medal of the National Institute of Arts and Letters—among other awards. Additionally, she was made Chevalier de l'ordre des Lettres in France and received a Gold Medal Award for Lifetime Achievement from the Poetry Society. Her notebooks constitute a unique and extraordinary resource for understanding the composition, experience, and intellectual alertness of a brilliant thinker and poet to a very broad range of popular, mundane, intellectual, artistic, and historically significant events of her time.

The MMDA is making digital reproductions and transcriptions of Moore's notebooks readily accessible to scholarly, classroom, and non-academic readers for the first time. The transcriptions are supported by annotations contextualizing Moore's writing and life, including citations to the original source texts she invokes, and an image-text linking feature that makes it easy to move back and forth between the facsimile and the transcription. The digital editions of the notebooks are supported by a growing collection of related materials, such as indexes, a glossary, an interactive timeline of Moore's life and publications, searchable reproductions of the now hard-to-find Marianne Moore Newsletter, and faceted text and image search tools currently under development. This site will, we hope, revolutionize criticism on this significant poet; contribute to popular understanding of the modernist period's history and culture; and develop new tools for the digital editing and publication of handwritten materials.

Moore's notebooks offer extraordinary challenges for editors and digital designers because they include multiple genres, images (Moore frequently sketched objects that interested her), genetic layers of text (evidence of Moore's later editing of and additions to earlier notes, sometimes with different writing implements), and references to decades' worth of popular and academic source materials. Fully edited and annotated, Moore's notebooks suggest the deep cultural genesis of her poems: Moore's notebooks constitute not nearly finished drafts of poems and essays (that is, pre-publication texts) but, instead, a rich and varied collection of notes and compilations from which Moore drew the materials that went into her drafts and published work. Part commonplace book, part scrapbook, part sketchbook, part diary, each of Moore's notebooks offers an extraordinary window into the eclectic print and visual culture of the twentieth century and the ways in which one of America's most innovative poets responded to her world.

The MMDA addresses the special challenges and opportunities presented by these documents through a customized version of the Edition Visualization Technology (EVT), an open-source software project directed by Roberto Rosselli Del Turco (Del Turco et al, 2014-15). We have extended the tool significantly to allow for the transcription and facsimile coordinates to be edited directly in the browser, then saved automatically to underlying TEI XML files according to our project's customized TEI schema (Our viewer/editor tool was developed on the basis of the EVT version 1.0, which utilizes XSLT, CSS, and Javascript to generate its viewer. The EVT version 2 recently released utilizes a different method that removes the need for XSLT and relies on converting the TEI XML documents into JSON, which we are considering as an alternative for future development. See the EVT blog for more information about the proposed changes for version 2). This has significantly extended the utility of the EVT, and we will continue to add more advanced editing features for modern manuscript collections. This combination of editing and publishing capacity makes the editing of these manuscripts accessible to any of our registered team members directly in the browser, allowing us to address the needs of the site's editorial team and its users through the development of an integrated web platform, which we are continuing to develop along with manuscript editing tools being developed at the University at Buffalo in collaboration with the Computer Science and Engineering and the Center for Unified Biometrics and Sensors (CUBS) at the University at Buffalo. Our partners in CUBS are also developing automated handwriting recognition software that we are integrating into our platform. We aim to create a complete editorial workflow solution for transcription and annotation of manuscript collections. This is possible through the use of the HubZero 2.0 platform, based on Joomla, which to our knowledge has not previously been used for a digital humanities project.

The *MMDA* is making a vast quantity of unpublished writing by a major American poet and cultural figure freely available and easily accessible for the first time. Such access will have extraordinary impact. It is already transforming Moore studies and, as the number of publications on the site increases, it will contribute significantly to historical, cultural, and literary studies of modernism and of twentieth-century women's lives. The specialized tools and platforms for collaboration on digital manuscript editions we are developing will, we hope, serve as useful models for making manuscript collections more accessible to scholars and editors in collaboration.

Bibliography

Del Turco, R.R., Buomprisco, G., Di Pietro, C., Kenny, J., Masotti, R., and Pugliese, J. (2014-15), "Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions", *Journal of the Text Encoding Initiative*, Issue 8 (December 2014 - December 2015). URL : http://jtei.revues.org/1077 ; DOI : 10.4000/jtei.1077

Leavell, L. (2014) *Holding On Upside Down: The Life and Work of Marianne Moore* (Farrar, Straus, Giroux)

Marianne Moore Digital Archive. (2015) University of Buffalo and the Center for Unified Biometrics and Sensors. www.moorearchive.org.

Miller, C. (2005) *Cultures of Modernism: Marianne Moore, Mina Loy, and Else Lasker-Schüler* (University of Michigan Press, 2005)

Miller, C. *(1995) Marianne Moore: Questions of Authority*. Harvard UP, 1995.

Opening up the Oxford English Dictionary: what an enhanced legacy dataset can tell us about language, lexicography, and literature

David-Antoine Williams
david.williams@utwaterloo.ca
St Jerome's University in the University of Waterloo
Canada

Overview

This paper will discuss recent research carried out in the context of an OMRI-funded project: "The Life of Words: Poetry and the OED." It will discuss the processes and methods developed for enhancing and manipulating a large legacy dataset—the Second Edition of the Oxford English Dictionary—and will present analyses and applications pertaining to lexicography, lexicology, and traditional literary studies. Styled as an "opening up" of latent information in a previously closed system, and with implications in digital humanities, linguistics, lexicography, lexicology, literary criticism, and poetics, the paper will address the conference theme of "Access" in the sense of tapping into knowledge that has previously been inaccessible.

Context

The Oxford English Dictionary (OED) is widely considered to be the greatest philological and lexicographical achievement in English. As the first fascicles of the OED were being prepared, editor James Murray professed his dedication to "The perfection of the Dictionary in its data" (1880: 129, orig. emph.). The "data" of the work is its 2.43 million quotations, a significant portion of them from poetic and other literary texts, which both shape and illustrate the various sense definitions of roughly 600,000 English words and word forms. Conversely, since its publication, poets have relied on the OED to guide their deployments and arrangements of English words in poems. This reciprocal intertextuality has led to two striking facts which have yet to be fully explored: 1) that the OED's definitions of English words depend to a significant degree on poetic language, which is striking because by any standard account, poetic usage tends away from the denotative or definitional and towards the connotative and metaphorical; and 2) that much English poetry of the last hundred years contains a philological, etymological, and lexicographical dimension, informed by the OED.

Although the Second Edition of the OED (Murray et al., 1989) was among the first large reference works to be prepared for public and academic communities in digitized, marked-up form, and despite the current and ongoing revision of the dictionary (Simpson et. al., 2000-), no version has ever been marked-up with additional metadata. Information we might expect to find in a modern text dataset, such as author gender, genre of quotation, and type of publication, is not included in the OED. Attempts to incorporate such information into studies of linguistic, literary and cultural questions have until now have therefore been limited to "case-study" or "sampling" methodologies.

"The Life of Words: Poetry and the OED" addresses this by working directly with legacy versions of the electronic OED, enhancing these with appropriate metadata about the quotation evidence. With the enhanced dataset, alongside modern large text corpora, we then generate quantitative and qualitative assessments in two broad fields of inquiry: 1) What has been the influence of poetry on the English language's most comprehensive lexicographical work? and 2) What influence has the OED had on English-language poetry? To take a modern turn on Murray, our concern now is the perfection of the dictionary in its data, metadata, and comparative data.

Paper Outline

The preamble to this paper will briefly introduce the project, discussing its background, methods employed, and current stage of development, and offering some observations regarding the use of dictionaries in general, the OED in particular, and specifically the "opened-up", marked-up and directly accessible enhanced OED, as "evidence" for interpretation in a number of scholarly domains, a methodological topic which has received recent attention in dictionary studies, literary studies, and linguistics (e.g. Coleman 2013a, 2013b; Coleman and Ogilvie 2009; Hoffman 2004). The bulk of the presentation will be devoted to an exploration of the enhanced OED, demonstrating some top-level findings relevant to current scholarship in three fields. In the history of lexicography, recently there has been much discussion on the interpretation of OED quotation evidence as a complex indicator of both the prestige of certain kinds of writing over time, and the particular judgements, biases, and practices of the nineteenth-century philologists who compiled the First Edition (1884-1928) (Ogilvie 2013; Brewer 2012, 2010, 2009, 2007; Considine 2009; Mugglestone 2005, 2000; Willinsky 1994). In the first main part of the paper, I give an overall quantitative assessment of the generic make-up of OED quotations, comparing the First and Second Editions, and discuss the implications of this for literary and for cultural history. Next, at the intersection of lexicology and literary studies, I offer a re-assessment of claims surrounding the linguistic inventiveness of canonical authors such as Shakespeare and Milton (Goodland 2011, 2010; Brewer 2013, Crystal

2000; Gray 1989; Schafer 1980) based on benchmarks for the period and genre of their various works. Finally, in the realm of literary criticism, I demonstrate a number of ways that information embedded in the OED can be harnessed to detect literary tropes such as allusion and etymological wordplay, either in poems that have been directly influenced by the OED, and those for which the evidence is less conclusive.

Bibliography

Brewer, C. (2013). "Shakespeare, word-coining, and the OED" in Shakespeare Survey 65: 345-57.

Brewer, C. (2012). "Happy Copiousness? OED's Recording of Female Authors of the Eighteenth Century" in Review of English Studies 63.258: 86-117.

Brewer, C. (2010). "The Use of Literary Quotations in the OED", Review of English Studies 61: 93-125.

Brewer, C. (2009). "The OED as 'literary instrument': its treatment past and present of the vocabulary of Virginia Woolf" Notes & Queries 56: 430-44.

Brewer, C. (2007). Treasure-House of the Language: The Living OED. New Haven: Yale University Press.

Coleman, J. (2013a). "Using Dictionary Evidence to Evaluate Authors' Lexis: John Bunyan and the Oxford English Dictionary" in Dictionaries The Journal of the Dictionary Society of North America 34: 66-100.

Coleman, J. (2013b). "Forum: Using OED Evidence" in Dictionaries: The Journal of the Dictionary Society of North America 34: 1-9.

Coleman, J. and Ogilvie, S. (2009). "Forensic Dictionary Analysis: Principles and Practice" in International Journal of Lexicography 22.1: 1-22.

Considine, J.(2009). "Literary classics in OED quotation evidence" in Review of English Studies 60.246: 620-638.

Crystal, D. (2000). "Investigating Nonceness: Lexical Innovation and Lexicographical Coverage" in R. Robert Boenig and K. Davis, eds, Manuscript, Narrative and Lexicon: Essays on Literary and Cultural Transmission in Honor of Whitney F. Bolton. Lewisburg: Bucknell University Press,: 218-31.

Goodland, G. (2011). " 'Strange deliveries': Contextualizing Shakespeare's first citations in the OED" in Mireille Ravassat and Jonathan Culpeper, eds, Stylistics and Shakespeare's Language: Trans disciplinary Approaches (London: Continuum,): 8-33.

Goodland, G. (2010). "The OED and 'single-use' words". http://ora.ouls.ox.ac.uk/objects/uuid%3A99d462ea-be60-4b60-b7a2-6259a862c500/datastreams/ATTACHMENT03.

Hoffman, S. (2004). "Using the OED Quotations Database as a Corpus—a linguistic appraisal." ICAME Journal, 28: 17–30.

Mugglestone, L. (2005). Lost for Words: The Hidden History of the Oxford English Dictionary. New Haven: Yale University Press.

Mugglestone, L. (2000). Lexicography and the OED: Pioneers in the Untrodden Forest. Oxford: Oxford University Press.

Murray, J. A. H. et. al., eds. (1989). Oxford English Dictionary. 2nd edn, compiled by J. A. Simpson and E. S. C. Weiner, 20 vols. Oxford: Oxford University Press.

Murray, J. A. H. (1880). 'The President's Annual Address for 1880', in Transactions of the Philological Society, 1880–1881. London: Trubner.

Ogilvie, S.(2013). Words of the World: A Global History of the Oxford English Dictionary. Cambridge: Cambridge University Press.

Schäfer, J. (1980). Documentation in the O.E.D.: Shakespeare and Nashe as Test Cases. Oxford: Clarendon Press.

Simpson, J., Weiner, E. S. C. and Proffitt, M. (2000-). OED Online. 3rd edn, rev. J. A. Simpson et al.. Oxford: Oxford University Press.

Short Papers

Using Archival Texts to Create Network Graphs of Musicians in Early Modern Venice

Mollie Ables
mables@indiana.edu
Indiana University, United States of America

Tassie Gniady
ctgniady@iu.edu
Indiana University, United States of America

Kalani Craig
craigkl@indiana.edu
Indiana University, United States of America

Grace Thomas
ghthomas@umail.iu.edu
Indiana University, United States of America

Adam Hochstetter
adamhoch@umail.iu.edu
Indiana University, United States of America

Introduction

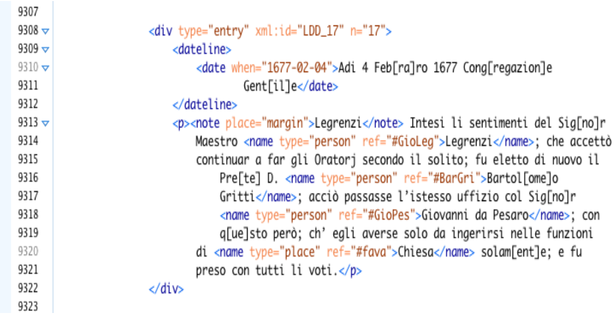
This project uses network graphs to depict musicians' careers in late seventeenth-century Venice. The current network graph, viewable on the Musicians in Venice web page, demonstrates relationships between musicians and the institutions that employed them. The graph is bimodal, with nodes representing musicians ("people" nodes) and institutions ("place" nodes). Archival texts are incorporated into the visualization, with transcriptions of records that indicate musicians' activity included in the node attributes. The entire project is text-based, with data derived from XML transcriptions of archival records in addition to assigned metadata. The current website is a proof of concept for a larger project that would demonstrate ways of displaying text as part of the network graph and using texts in creating network visualizations.

The current graph focuses on the career of the composer Giovanni Legrenzi. Legrenzi worked for several prominent Venetian institutions from the early 1670s to his death in 1690 and ultimately was appointed to the most prestigious musical posts in the city. The musicians he worked with also served multiple institutions, either simultaneously or in succession, and their relationships with other musicians, patrons, and administrators often facilitated their movement between institutions. Legrenzi's Venetian career presents an excellent case study of these musical connections in late seventeenth-century Venice, and studying these connections demonstrates how networks of musicians functioned in this time period. This study also provides a representative sample of how network visualization effectively demonstrates patterns in musicians' careers in Venice.

The texts in the network graph include transcribed administrative documents, mostly unpublished, from the Venetian institutions where Legrenzi was active between 1670 and 1690. These are primarily payment, hiring, and termination records, which document the activity of the musicians employed by or affiliated with these institutions. Several generations of musicologists have used these documents to identify where different musicians were working, when they were employed, and in what capacity (Bonta, 1964; Moore, 1981; Termini, 1981; and Selfridge-Field, 1994). In this sense, this treatment of the documents is standard to the field, but applies DH methodology and processes to create visual representations the data. This provides a new perspective on the sources. For instance, grouping the "person" nodes by centrality results in multiple sub-groups, and proximity among these nodes demonstrates shared institutions and implied communities. (Hanneman and Riddle, 2005: Chapter 10) (Mary Russell Mitford's text-based network analysis of Robert Southey's *Thalaba the Destroyer* provides an excellent example of this. In addition to demonstrating the behavior of agile, well-connected nodes, the graph displays modules of nodes as units of analysis (Wasserman and Faust, 1994: 4).

Methodology

The transcriptions of the original documents were encoded in XML with tags and attributes determined by the Text Encoding Initiative (TEI). I used the XML markup to tag the information I wanted to include in my data while retaining the complete original text. In general, I tagged the name of the musician, the name of the institution with which they were connected, and the date and location of the connection.



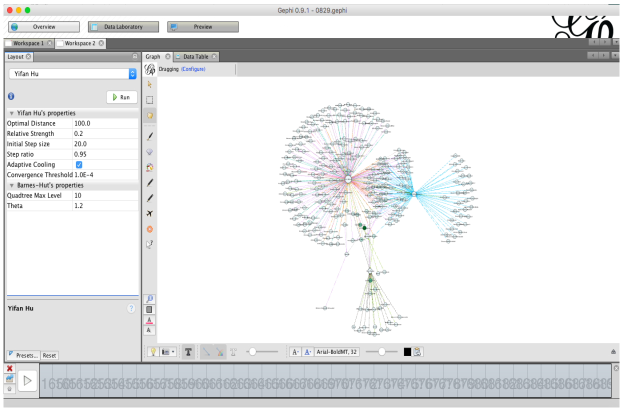
```
        <div type="entry" xml:id="LDD_17" n="17">
            <dateline>
                <date when="1677-02-04">Adi 4 Feb[ra]ro 1677 Cong[regazion]e
                        Gent[il]e</date>
            </dateline>
            <p><note place="margin">Legrenzi</note> Intesi li sentimenti del Sig[no]r
                Maestro <name type="person" ref="#GioLeg">Legrenzi</name>; che accettò di
                continuar a far gli Oratorj secondo il solito; fu eletto di nuovo il
                    Pre[te] D. <name type="person" ref="#BarGri">Bartol[ome]o
                    Gritti</name>; acciò passasse l'istesso uffizio col Sig[no]r
                    <name type="person" ref="#GioPes">Giovanni da Pesaro</name>; con
                    q[ue]sto però; ch' egli averse solo da ingerirsi nelle funzioni
                di <name type="place" ref="#fava">Chiesa</name> solam[ent]e; e fu
                preso con tutti li voti.</p>
        </div>
```

XML markup of primary document text

I then extracted the information for the network from the XML document using an XSLT. For this project, the stylesheet extracted the data assembled in the person index of the XML document and combined it with the entire text that documented each event. The resultant document was a CSV file of raw data that had to be separated into sets of nodes and edges to be read by the visualization platform.

The data visualization for this project was generated using Gephi, which is an open-source platform for exploratory data analysis. For this project, I generated the network graph in Gephi by importing two CSV files: one with the nodes and their attributes and ID numbers, and one with the edges and the source and target data. The online version of the network graph includes all the text associated with a "person" node in the network as a node attribute. When the user selects a person node in the online version, it displays the transcriptions for every document mentioning that person in the Information Pane. (I exported the data using the SigmaExporter plugin for Gephi. Version 0.9.0 from Gephi Thirdparties Plugins. The code is available on Github.)

Displaying all the text associated with a person node presented a challenge as the XSLT exported each connection between a musician and an institution as a new line in the CSV. For instance, if a musician was hired by a church in 1674, received a raise in 1678, and retired in 1686, there would be three lines in the CSV. When imported into Gephi, this would create three connecting edges between the person node and the place node. This created a misleading network graph, as the number of documented connections between people and places depended on the nature and availability of the records. To combine all the texts associated with the events for each person in the index, I concatenated the raw data using a PHP script and saved the results as a TSV file that could be imported into Gephi. As a result, the entire context of the decrees and decisions surrounding that musician's activity appears in the Information Pane, which also displays and links to the institutions associated with that musician. The PHP script also included HTML formatting to better distinguish the entries from one another in the visualization. Once my nodes and edges were imported into Gephi, I used a layout that grouped the nodes by degree of centrality (Algorithm in Gephi based on Blondel, Guillaume, Lambiotte, and Lefebvre, 2008. Resolution in Gephi based Lambiotte, Delvenne, and Barahona, 2009).



Yifan Hu

Yifan Hu's properties

Optimal Distance	100.0
Relative Strength	0.2
Initial Step size	20.0
Step ratio	0.95
Adaptive Cooling	☑
Convergence Threshold	1.0E-4

Barnes-Hut's properties

Quadtree Max Level	10
Theta	1.2

Network graph in Gephi

Next steps

The current graph is the foundation of a more extensive project that uses musicians and musical activity in Early Modern Venice to benefit scholars working with TEI in similar for-hire environments. In these environments, hubs of activity define the relationship between individual practitioners in the historical equivalent of a sharing economy. In the long-term, I propose an online data repository and network graph of musicians working in seventeenth-century Venice that would eventually provide a roadmap for scholars embedded in the TEI transcription model, but with an interest in an automated process for applying TEI text analysis to network analysis.

In addition to the bimodal graph that highlights relationships between musicians and their employing institutions, I also hope to create a unimodal graph documenting relationships among the musicians. Creating the unimodal network and a supporting web application will be a challenge as relationships between individuals can be complicated and dynamic. This will take considerable work in effectively diagramming entity relationships and building web applications that go beyond the "out of the box" functionality of the Gephi export plugin. The website for the Six Degrees of Francis Bacon project provides an excellent example of a unimodal graph; here the relationships are defined by circumstances (such as "father to" or "colleague of") or actions (such as "met" or "wrote to"). Linked Jazz, which uses documents as node attributes to demonstrate relationships and meaningful connections between jazz musicians in an interactive network graph, is also an excellent example. Both sites feature custom-designed entity relationships and interactive features in their web design that I want to emulate in my own network graph.

For the unimodal version, I would expand the data beyond Legrenzi's Venetian career, increasing the time frame and the number of institutions represented in the graph, and include different kinds of sources, such as periodicals, correspondence, and notarial records. Developing relationship typologies is such a crucial component to the project. This will require creating custom node metadata based on the information from the primary sources and agile development as the number and variety documents expand. The result will be a project that can better serve other scholars of Venetian music and culture.

Bibliography

Bonta, S. (1964). "The Church Sonatas of Giovanni Legrenzi." Ph.D. diss., Harvard University.

Hanneman, R. and Riddle, M. (2005). *Introduction to Social Network Methods*. 2005. http://www.faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html

Moore, J. H. (1981). *Vespers at St. Mark's: Music of Alessandro Grandi, Giovanni Rovetta, and Francesco Cavalli*. Ann Arbor, Michigan: UMI Research Press.

Selfridge-Field, E. (1994). *Venetian Instrumental Music from Gabrieli to Vivaldi*. New York: Dover.

Termini, O. (1981). "Singers at San Marco in Venice: The Competition between Church and Theatre (c1675 - c1725)." *Royal Musical Association Research Chronicle* 17: 65-96

Wasserman, S. and Faust, K. (1994). *Social Network Analysis:*

Methods and Applications. Cambridge: Cambridge University Press.

How does Google Cultural Institute Affect the Collections of Museums? The Case of Turkey

Sümeyye Akça
sumeyyeakca@hacettepe.edu.tr
Hacettepe University, Turkey

Introduction

In recent years, progress on information and communication technology (ICT) has affected the management of cultural heritage. While ICT plays a key role in the accessibility and informed experiences of the public, it is also becoming more apparent in creating participatory platforms for people who manage and enjoy cultural heritage (Lekakis and Chrysanthi, 2011). Thus cultural memory institutions use this technology for creating digital content about their collections and making this content available from all over the world. Increasing the value of collections using sophisticated and innovative new media also affects the economic development of countries and provides more integrated awareness about cultural identity and cross-cultural communication (Brizard, Derde, Silberman, 2007).

ICT has provided a wide range of tools for cultural heritage management. Implementing these new tools also provides many advantages for both memory institutions and users. Memory institutions can manage their collections more easily with support for image processing, advanced publishing systems, and etc. Furthermore, while open access to public involvement in cultural heritage objects improves the public awareness and sense of belonging to the society, it also develops the content of the objects with crowdsourcing from all over the world. In order to benefit from these advantages of new media tools, cultural memory institutions have opened their collections to the global public regardless of geographic location. Thus, memory institutions also benefit from increasing their promotion and advertising among other countries and institutions (Myat, 2012).

One of these new media platforms is Google Cultural Institute, which has been applied to the cultural heritage area in recent years, especially in museums. The purpose of this institution is to provide broad public awareness, use and augmented access online for cultural heritage objects. It offers a range of tools that make it easy for memory institutes to put the collections online. With the online exhibitions, memory institutions can create stories about their collections. People can reach these collections all around the world using a wide range of platforms, and can share the images they like with friends using their social media accounts (Google Cultural Institute, 2016). In his 2016 TED talk, the head of Google Cultural Institute, Amit Sood says "The world is filled with incredible objects and rich cultural heritage. And when we get access to them, we are blown away, we fall in love. But most of the time, the world's population is living without real access to arts and culture." This new platform allows users to experience the world's cultural heritage objects, and to add comments about them as well.

Method

There are ten museums in Turkey putting their collections on Google Cultural Institute. All of these ten museums are private. They are as follows:

1. Salt
2. Rezan Has Museum
3. Elgiz Museum
4. Sakıp Sabancı Museum
5. Borusan Contemporary
6. Murat Istanbul Festival
7. Masumiyet Museum
8. Istanbul Museum of Modern Art
9. Istanbul Research Institute
10. Pera Museum

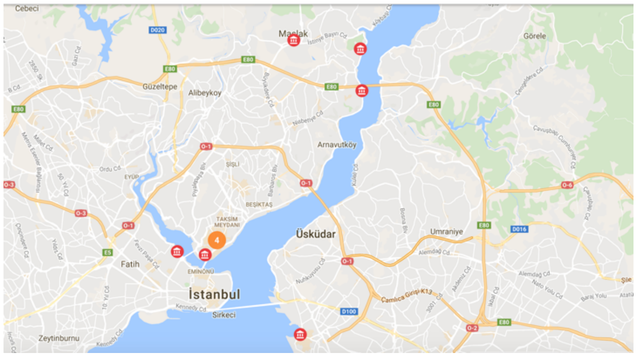


Figure 1. The partners of Google Cultural Institute from Turkey (accessed on Google Cultural Institute, 2016).

In this study, the data about visibility and access of these ten museums is analyzed looking at the previous and next data of the collections that are added to this online platform. For detailed analysis of this issue, interviews were conducted with curators and managers of these ten museums regarding their observations of this new platform. The web activities of the museums -before and after- are also visualized. In addition, the data about the social media usage of these museums and its effect on increasing visibility and access of the collections are discussed. To understand the data in this study, these research questions are answered.

1. How does this online platform affect the visibility of these collections?
2. Does it affect the diversity of visitors being national or international?

3. What are the advantages and disadvantages of this new platform?

Bibliography

Brizard, T., Derde, W., & Silberman, N. (2007). Basic Guidelines for Cultural Heritage Professionals in the Use of Information Technologies. How can ICT support cultural heritage. Accessed on 31.10.2016 http://www.enamecenter.org/files/documents/Know-how%20book%20on%20Cultural%20Heritage%20and%20ICT.pdf.

Google Cultural Institute. (2016). Accessed on 31.10.2016 https://www.google.com/culturalinstitute/beta/partner?tab=map

Lekakis, S. and Chrysanthi, A. (2011). Sustainable heritage management...for whom? A critique on contemporary economics of culture & the use of Information and Communication Technologies towards a symbiotic management strategy. *The 5th Biennial HO PhD Symposium on Contemporary Greece and Cyprus* (London, UK 2-3, June 2011). Accessed on 31.10.2016 http://www.lse.ac.uk/europeanInstitute/research/hellenicObservatory/CMS%20pdf/Events/2011-5th%20PhD%20Symposium/Lekakis.pdf

Myat, A. (2012). Social media technology in cultural heritage. Accessed on 31.10.2016 http://blogs.ntu.edu.sg/h6716-heritage/.

Sood, A. (2016) Every piece of art you've ever wanted to see — up close and searchable. [Ted Talk]. February 2016. Accessed on 31.10.2016 https://www.ted.com/talks/amit_sood_every_piece_of_art_you_ve_ever_wanted_to_see_up_close_and_searchable/transcript?language=en

Digital *Hansard*: Politics and the Uncivil

Marc Alexander
marc.alexander@glasgow.ac.uk
University of Glasgow

Andrew Struan
andrew.struan@glasgow.ac.uk
University of Glasgow

Summary

This short paper uses the recently-completed Hansard Corpus to show the patterning of attitudes expressed by the British Parliament about things considered to be ‘uncivilized’ across the last two centuries. It starts from the lexical resource of the Historical Thesaurus of English to gain an overview of the lexicalisation of the concept ‘uncivilized’ and uses this digital data demonstrates a substantial shift (from foreign to domestic) in who Parliament considers to be uncivil.

Introduction

The ways in which the British have discussed ‘uncivilized’ peoples which travellers have encountered throughout the history of English gives a key insight into how people in the past have identified and classified the world around them. This paper uses data from the *Hansard Corpus 1803-2003* (Alexander and Davies, 2015-) alongside the *Historical Thesaurus of English* (Kay et al, 2015–) to analyse the evolution of how the English-speaking people have thought of those who they think uncivil in five different sense-families — as animals, as ill-formed people, as strange-speaking outsiders, as savages, and finally as innocents awaiting enlightenment. Only these large digital data sources can show us the patterning of who and what the British Parliament have considered to be barbarous across time.

Data

This analysis became possible following the completion of the *Historical Thesaurus of English* (HT) in 2009 and the semantically-tagged *Hansard Corpus 1803-2005* in 2015, both of which are currently directed by Alexander and were created by teams of scholars at the University of Glasgow.

The HT is a database of all the recorded words in the history of English arranged according to their meaning; one of the world’s oldest digital humanities projects, and in progress for over 50 years, the HT database (stored on media from punch cards to tape to diskettes to networked storage to the Web) allows us an unparalleled resource for analysing the history of English. The *Hansard Corpus 1803-2005*, completed in 2015, is a digital corpus of speeches in the British Parliament between those dates, consisting of 1.6bn words across 7.6m speeches. Its contents were semantically tagged in the 2014-15 SAMUELS project (The SAMUELS Consortium, 2015) with disambiguated meaning codes from the HT, making it possible to search for semantic categories rather than words, as we do below.

The Uncivil

The category of *Civilization* in the HT gives us an indication of a non-typical pattern in the number of words available to describe a given concept (in English, categories normally grow throughout time) in the words referring to *uncivilized* and a *lack of civilization*, as Figure 1 shows.

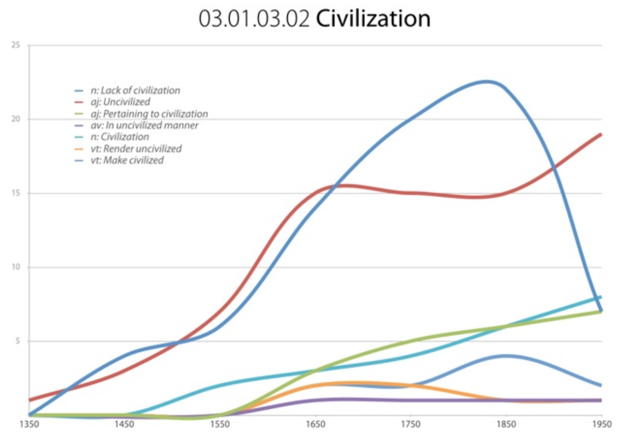
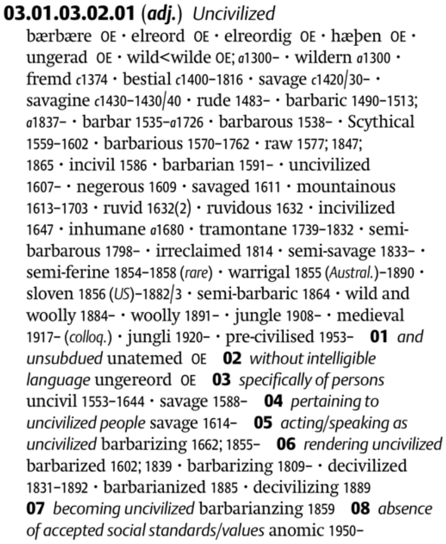


Figure 1: The size of each subcategory of *Civilization* in the HT

While the size of the *uncivilized* adjective category rises in the latter 20th century, there is a substantial fall at the same time in the size of the *lack of civilization* noun category, which we argue is connected to the shift in who has been considered to be uncivil (see below). In addition, of the 42 words in the *uncivilized* category in the HT (see Figure 2), the vast majority follow a particular path of lexicalization which we describe below, with new terms reflecting the shifting conceptualization of the uncivil throughout the times at which they were coined.



03.01.03.02.01 (***adj.***) *Uncivilized*
bærbære OE · elreord OE · elreordig OE · hæþen OE · ungerad OE · wild<wilde OE; *a*1300– · wildern *a*1300 · fremd *c*1374 · bestial *c*1400–1816 · savage *c*1420/30– · savagine *c*1430–1430/40 · rude 1483– · barbaric 1490–1513; *a*1837– · barbar 1535–*a*1726 · barbarous 1538– · Scythical 1559–1602 · barbarious 1570–1762 · raw 1577; 1847; 1865 · incivil 1586 · barbarian 1591– · uncivilized 1607– · negerous 1609 · savaged 1611 · mountainous 1613–1703 · ruvid 1632(2) · ruvidous 1632 · incivilized 1647 · inhumane *a*1680 · tramontane 1739–1832 · semi-barbarous 1798– · irreclaimed 1814 · semi-savage 1833– · semi-ferine 1854–1858 (*rare*) · warrigal 1855 (*Austral.*)–1890 · sloven 1856 (*US*)–1882/3 · semi-barbaric 1864 · wild and woolly 1884– · woolly 1891– · jungle 1908– · medieval 1917– (*colloq.*) · jungli 1920– · pre-civilised 1953– **01** *and unsubdued* unatemed OE **02** *without intelligible language* ungereord OE **03** *specifically of persons* uncivil 1553–1644 · savage 1588– **04** *pertaining to uncivilized people* savage 1614– **05** *acting/speaking as uncivilized* barbarizing 1662; 1855– **06** *rendering uncivilized* barbarized 1602; 1839 · barbarizing 1809– · decivilized 1831–1892 · barbarianized 1885 · decivilizing 1889 **07** *becoming uncivilized* barbarianzing 1859 **08** *absence of accepted social standards/values* anomic 1950–

Figure 2: 'Uncivilized' in the HT, taken from p.1235 of the print edition.

Thus far this sort of analysis has been slow-paced and difficult to undertake. However, with the tagging in the *Hansard Corpus 1803-2005* we can investigate this sort of semantic and conceptual change in a much more rapid fashion by honing in on uses of these meanings in context across time.

Parliament

There are five families of meaning into which the words above can be categorised, as outlined above. In a past article (Alexander and Struan, 2013), we assembled some evidence for this from the history of English in a non-systematic fashion. For this short paper, we instead account for all the evidence from the *Hansard Corpus* — over 2,000 uses of the semantic category — in order to trace across recorded Parliamentary history the shifts in the cultural, political and social attitudes towards the 'uncivilized'. This shows a substantial change in the picture which differs from the simpler five-family view of the sense evolution of *uncivil* we described in that earlier article.

Our first change to discuss is the shift, shown below, from the uncivil primarily being foreigners in the 1800s to being domestic persons in the 1900s onwards.

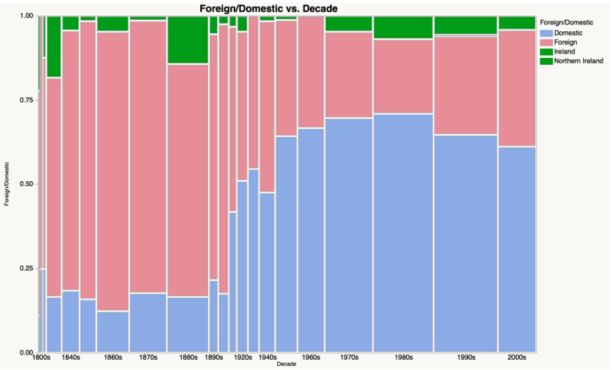


Figure 3: The proportion of uses of *uncivil* words to refer to foreign or domestic persons (thickness of bars reflects the changing amounts of text in *Hansard* in those decades); note that the status of Ireland and Northern Ireland is contested with regards to the foreign/domestic status, and so has been represented separately here

This is reflected in the changing discourse surrounding *barbaric* and *uncivil* things, where a majority of 20th century uses refer to barbaric practices and actions rather than persons:

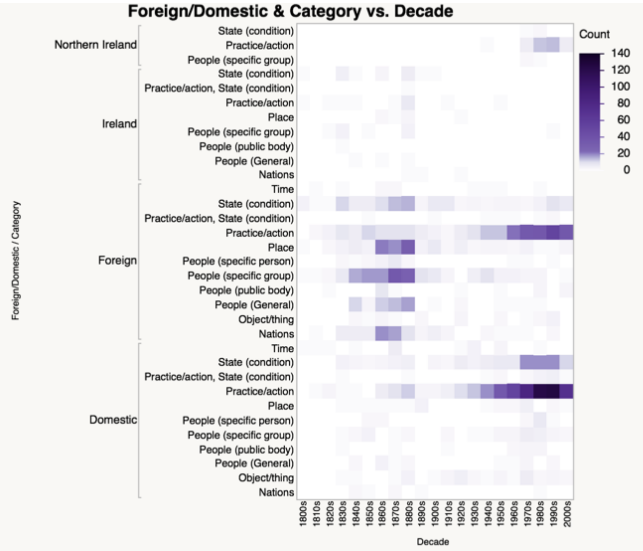


Figure 4: A heatmap of the entities (people, states, practices) considered uncivil by Parliament in the data, separated by whether the entities are foreign or domestic

Through four other graphs, we further report on the distribution of uncivil references across the globe and between the two Houses of Parliament. We also show the changes in the five evolutionary sense-families we outline above, which is key to the foreign/domestic shift we describe.

Some quotes from the corpus can briefly illustrate these changes, which here are aimed at a general body of persons, or a country:

> Mr Charles Adderley, House of Commons 21 February 1865: *'...to discharge what Lord Grey described as the singular office of dispensing rude laws among uncivilized tribes.'*
>
> Earl of Carnarvon, House of Lords 12 May 1874, on India: *'But a central government is not enough. In barbarous times and in uncivilized countries, roads are the first condition of improvement; and here it will be our first duty to open and secure the maintenance of roads and trade-paths.'*
>
> Mr Richard Cherry (Attorney-General for Ireland), House of Commons 20 March 1908: *'I never said that the people of Ireland were West African savages.'*
>
> Lord Hylton, House of Lords 18 April 1995: *'We can now see that in dealing with Russia we are dealing with a semi-barbarous state and a society that only knew a measure of democracy for a few years before the First World War.'*
>
> Mr Andrew Robathan, House of Commons 1 November 2001, on the pending invasion of Iraq: *'We should not allow a barbaric, mediaeval [sic] regime to succeed or last. We certainly do not want to go back to civil war.'*

As a result, we can show empirically the shift over two centuries in the ways which Members of Parliament described uncivil or barbaric entities, from foreign people or places to domestic practices. We conclude by arguing that this is the result of increased oppositionality being shown in the digital Parliamentary record, and so in this short paper we combine 'big picture' graphs of large-scale data analysis with more focused examples from the corpus record.

Bibliography

Adamson, S., Allan, K., Andrade, S., Arac, J., Davis, J., Durant, A., Durkin, P., Heath, S., MacCabe, C., Mehl, S., Robertson, K., and Yanacek, H. (2016). The Keywords Project. University of Pittsburgh. (http://keywords.pitt.edu/index.html)

Alexander, M. and Davies, M. (2015–) *Hansard Corpus 1803-2005*. Available online at http://www.hansard-corpus.org.

Alexander, M. and Struan, A. 2013. 'In countries so unciviliz'd as those?': The Language of Incivility and the British Experience of the World. In Martin Farr & Xavier Guégan (eds.) *Experiencing Imperialism: The British Abroad since the Eighteenth Century*, volume 2. London: Palgrave Macmillan.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., **and Alexander, M.** (eds.). 2015–. *The Historical Thesaurus of English*, version 4.2. Glasgow: University of Glasgow. http://www.glasgow.ac.uk/thesaurus

The SAMUELS Consortium (2015). *The SAMUELS Project*. United Kingdom AHRC and ESRC. http://www.glasgow.ac.uk/samuels.

Williams, R. (2014). *Keywords: A vocabulary of culture and society*. Oxford University Press

Reconstructing Readerly Attention: Citational Practices and the Canon, 1789–2016

Mark Algee-Hewitt
mark.algee-hewitt@stanford.edu
Stanford University, United States of America

David McClure
dclure@stanford.edu
Stanford University, United States of America

Hannah Walser
walser@stanford.edu
Stanford University, United States of America

In its ability to extract feature sets, relate texts within an abstract space, and semantically parse groups of texts, computational textual analysis has functioned primarily as a formalist intervention into literary study. When practitioners venture outside of the formal features of the texts themselves, it is author or date that serves as the point of contact between the text and its wider context. And yet, texts offer a rich history of reception: as different interpretive communities (Fish 1980) receive and reinterpret novels, poems or plays, they recontextualize the literary object to suit the particular socio-cultural goals of their period or nationality. Lacking detailed accounts of reading practices at large scales, even traditional practitioners of literary history have been unable to reconstruct the history of reception of even the most historically canonical texts. In this project, we leverage the ability of Digital Humanities to recover, at least provisionally, a large-scale history of textual reception by exploring the patterns of citation that reveal the attention paid to specific texts across their history as readerly objects. How are certain canonical texts cited over time and what can the attention paid to different segments of texts with a rich reception history tell us about the reading or social practices of different historical periods? How do different groups of readers (particularly authors and critics) quote text differently as they make use of passages in their own writing? And how do specialists and non-specialists cite the same text differently? By exploring the locus of attention within a canonical text, both across groups of readers and across history, we provisionally reconstruct a historically and socially contingent map of a text's reception history.

As Piper and Algee-Hewitt have argued in "The Werther Effect" (Piper and Algee-Hewitt 2014), practices of citation, the embedding of the language of a text within other works, can reveal patterns of reception even within a single author's corpus. In this project, we expand this approach multi-dimensionally, identifying passages of canonical works quoted in other novels, in critical articles by specialists and non-field specialists, and in a larger undifferentiated corpus of text. The scale of our analysis enables us to identify what parts of a text have received the most writerly attention overall and how that attention has been shaped over time. By moving from semantics to passages, we switch our attention from the intangible metrics of semantic similarity, to specific, quotation-level instances of citation that demonstrate specific attention to identifiable parts of our target texts. Drawing on work in sequence alignment by David Smith et al (2013) and Richard So et al. (forthcoming), we will therefore be able to explore patterns of attention that have been paid to a text by identifiable groups of readers. We argue that these patterns of citationality serve as a proxy for the reception of a text: while necessarily limited to readers who themselves were authors (or critics), they nevertheless represent an important category of reception available to analysis.

To extract the quotations, we used Python's "difflib" module, wrapped up as a parallelized MPI program that runs on an HPC cluster. To compare any two individual texts - for example, when checking for passages from Hamlet inside of a novel from the Gale American Fiction corpus - the texts are first split into tokens and passed through a filter that removes a set of 200 stopwords. This speeds up the alignment algorithm (the high-frequency words that get pulled out make up a significant portion of the total words in any given text, producing shorter sequences) and also has the advantage of making the alignment process less sensitive to small changes in function words, which seem to get shuffled around or changed fairly frequently when a text is quoted. For example, a change from:

> And crook the pregnant hinges of the knee
> **Where** thrift may follow fawning
>
> to
>
> And crook the pregnant hinges of the knee
> **That** thrift may follow fawning

still gets picked up as a quotation, since the semantically significant words - crook, pregnant, hinges, knee, thrift, fawning - stay the same.

These filtered sequences of tokens then get passed through the alignment algorithm, which produces a set of matches, recoded in terms of their starting positions in each text and the length of the matching subsequence. To ensure that the matches represent actual quotations, we discarded matches shorter than 5 tokens (not counting stopwords), since alignments shorter than this include a fair number of false positives, generic word sequences that likely don't represent any kind of meaningful quotation or intertextuality - for example, many are numbers, things like "five hundred thousand." This gives us high "precision" - almost all of the alignments that are included in the final analysis represent legitimate quotations to the play - but it also drops down the "recall" somewhat, since some of the shorter alignments are, in fact, real quotes - things like "weighing delight and dole." We are currently evaluating a couple of strategies for identifying these alignments that are short but semantically "focused" enough that we can say with confidence that they should be included in the set of valid matches.

We use this method of sequence alignment to trace the quotations of five canonical texts with a rich citation history across four corpora. Our selected texts include Shakespeare's *Hamlet*, Milton's *Paradise Lost*, Dickens' *A Christmas Carol*, Caroll's *Alice in Wonderland* and Wordsworth's *Prelude*. Not only are all of these texts heavily quoted by critics, but, we argue, they have entered the literary and cultural consciousness of both Britain and America such that the passages that are cited by authors and critics reveal interpretive and readerly practices both across time and between different groups. For each text, we extract all of the citations from it that are five words (excluding stopwords) or longer, that occur in each of our four corpora: the full-text Hathi trust corpus, representing a massive sample of writing in the nineteenth and early twentieth centuries; a literature-specific corpus of 28,000 novels from England and America dating from 1789-2016; and two corpora of articles on literary studies, one a corpus of 10 journals of literary criticism and history (e.g. *PMLA*, *NLH*, *Critical Inquiry*) and one a corpus of 10 field-specific journals focused specifically on the authors represented in our group of canonical texts (e.g. *Shakespeare Quarterly*, *Milton Studies*, *Wordsworth Circle*).

Between these four corpora, we are able to differentiate the kind of attention paid to our primary canonical texts by four different groups of readers. Do novelists pay attention to different parts of a text than authors in general in the nineteenth and twentieth centuries? Do general literary critics quote different parts of *Hamlet* than Shakespeare specialists? And for each of these corpora, how does the citation map of each of our texts change over time?

For example, a citation map of *Hamlet* in our novel corpus revealed 1,693 quotations of five or more non-stopword tokens, which collectively cover about 25% of all words in the play. When we plot the frequency of citations across the narrative of the drama (broken into 500 bins), our method reveals the passages most quoted by novelists of the nineteenth and twentieth centuries (Figure 1). From this citation map, we can see that Hamlet's soliloquy ("to be or not to be") is among the top three passages cited by novelists; however, the quotation that clearly dominates the use of *Hamlet* by this group of readers comes from Act 5, Scene 2 “There’s a divinity that shapes our ends, / Rough-hew them how we will.—“

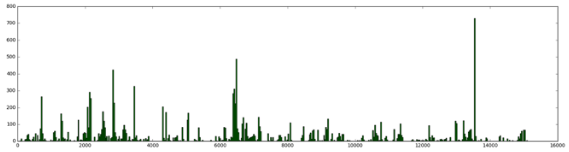


Figure 1 Numbers of citations of Hamlet in nineteenth and twentieth-century novels. Each passage is 1/500 of the text.

Although not among the most identifiable passages today, this quotation clearly had a resonance for the readers of the nineteenth and early twentieth centuries.

By comparing this map of citations across the narrative of Shakespeare’s plays to ones that are both drawn from our comparative corpora and periodized across the two centuries they represent, we are able to show how different passages gain and lose meaning across time and between kinds of reading. As we expand this to all five of our canonical texts read into all four of our corpora, we can shed light on how historically and disciplinarily specific practices of reading shaped the horizons of interpretation for specific works, and begin to reconstruct these reader-based practices in ways that are open and tractable to the Digital Humanities.

Bibliography

Fish, S. (1980) *Is there a text in this class?* Cambridge: Harvard UP.

Piper, A., and Algee-Hewitt, M. (2014). “The Werther Effect I: Goethe, objecthood and the handling of knowledge.” *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*. Ed. Matt Erlin and Lynn Tatlock. Rochester: Camden House. 155-184.

Smith, D., Cordell, R., Dillon, E. M. (2013). “Infectious texts: modeling text reuse in nineteenth-century newpapers.” *Proceedings of the IEEE International Conference on Big Data*. 86-94.

So, R. J.; Long, H, Yuancheng, Z. (forthcoming). “The Dark Code: Modeling White-Black Literary Relations, 1880-2000. *Forthcoming.*

Natural Language Processing for the Long Tail

David Banman
dbamman@berkeley.edu
UC Berkeley, United States of America

Natural language processing (NLP) is a research area that stands at the intersection of linguistics and computer science; its focus is the development of automatic methods that can reason about the internal structure of text. This includes **part-of-speech tagging** (which, for a sentence like *John ate the apple*, infers that *John* is a noun, and *ate* a verb), **syntactic parsing** (which infers that *John* is the syntactic subject of *ate*, and *the apple* its direct object), and **named entity recognition** (which infers that *John* is a PERSON, and that *apple* is not, for example, an ORGANIZATION of the same name). Beyond these core tasks, NLP also encompasses sentiment analysis, named entity linking, information extraction, and machine translation (among many other applications).

Over the past few years, NLP has become an increasingly important element in computational research in the humanities. Automatic part-of-speech taggers have been used to filter input in topic models (Jockers, 2013) and explore poetic enjambment (Houston, 2014). Syntactic parsers have been used to help select relevant context for concordances (Benner, 2014). Named entity recognizers have been used to map the attention given to various cities in American fiction (Wilkens, 2013) and to map toponyms in Joyce’s *Ulysses* (Derven et al., 2014) and Pelagios texts (Simon et al., 2014). The sequence tagging models behind many part-of-speech taggers have also been used for identifying genres in books (Underwood et al., 2013).

There is a substantial gap, however, between the quality of the NLP used by researchers in the humanities and the state of the art. Research in natural language processing has overwhelmingly focused much of its attention on English, and specifically on the domain of news (simply as a function of the availability of training data). The Penn Treebank (Marcus et al., 1993)—containing morphosyntactic annotations of the *Wall Street Journal*—has driven automatic parsing performance in English above 92% (Andor et al., 2016); part-of-speech tagging on this same data now yields accuracies over 97% (Søgaard, 2011). While a handful of other high-resource languages (German, French, Spanish, Japanese) have attained comparable performance on similar data (Hajič et al., 2009), many languages simply have too few resources (or none whatsoever) to train robust automatic tools. Even within English, out-of-domain performance of many NLP tasks—in which, for example, a syntactic parser trained on the *Wall Street Journal* is used to automatically label the syntax for *Paradise Lost*—is bleak. Figure 1 illustrates one sentence from *Paradise Lost* automatically tagged and parsed using a tool trained on the *Wall Street Journal*. Since this model is trained on newswire, it expects newswire as its input; errors in the part-of-speech assignment snowball to bigger errors in syntax.

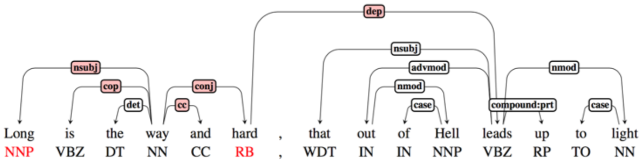
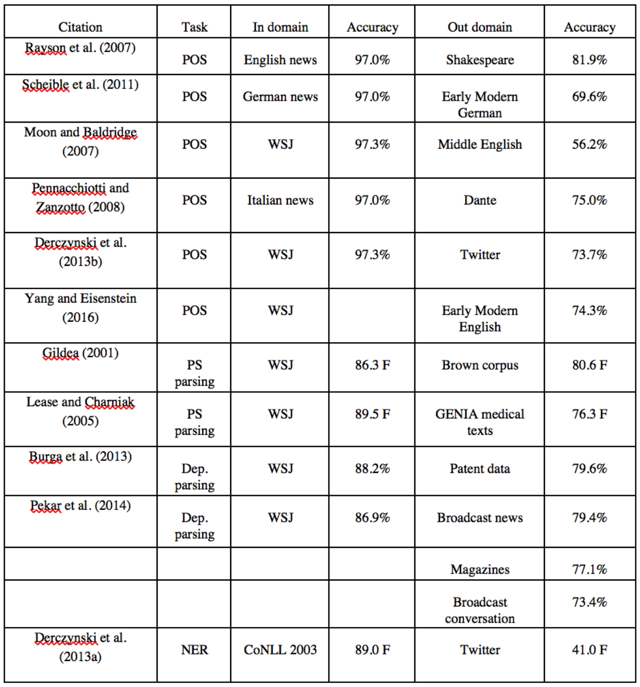


Figure 1: Parsers and part-of-speech taggers trained on the WSJ expect newswire syntax. Automatically parsed syntactic

dependency graph with part-of-speech tags for Long is the way and hard, that out of Hell leads up to light. Errors in part-of-speech tags and dependency arcs are shown in red. Part-of-speech errors snowball into major syntactic errors.

Table 1 provides a summary of recent research that has investigated the disparity between training data and test data for several NLP tasks (including part-of-speech tagging, syntactic parsing and named entity recognition). While many of these tools are trained on the same fixed corpora (comprised primarily of newswire), they suffer a dramatic drop in performance when used to analyze texts that come from a substantially different domain. Without any form of adaptation (such as normalizing spelling across time spans), the performance of an out-of-the-box part-of-speech tagger can, at worse, be half that of its performance on contemporary newswire. On average, differences in style amount to a drop in performance of approximately 10-20 absolute percentage points across tasks. These are substantial losses.



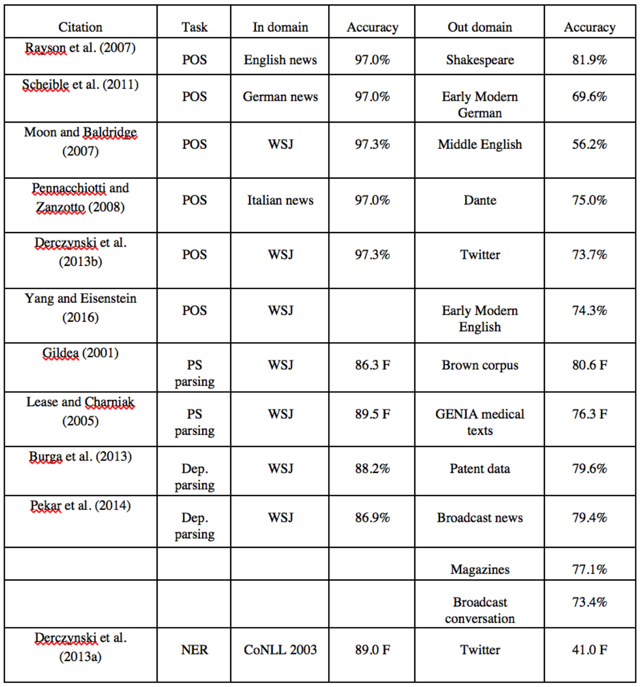
Citation	Task	In domain	Accuracy	Out domain	Accuracy
Rayson et al. (2007)	POS	English news	97.0%	Shakespeare	81.9%
Scheible et al. (2011)	POS	German news	97.0%	Early Modern German	69.6%
Moon and Baldridge (2007)	POS	WSJ	97.3%	Middle English	56.2%
Pennacchiotti and Zanzotto (2008)	POS	Italian news	97.0%	Dante	75.0%
Derczynski et al. (2013b)	POS	WSJ	97.3%	Twitter	73.7%
Yang and Eisenstein (2016)	POS	WSJ		Early Modern English	74.3%
Gildea (2001)	PS parsing	WSJ	86.3 F	Brown corpus	80.6 F
Lease and Charniak (2005)	PS parsing	WSJ	89.5 F	GENIA medical texts	76.3 F
Burga et al. (2013)	Dep. parsing	WSJ	88.2%	Patent data	79.6%
Pekar et al. (2014)	Dep. parsing	WSJ	86.9%	Broadcast news	79.4%
				Magazines	77.1%
				Broadcast conversation	73.4%
Derczynski et al. (2013a)	NER	CoNLL 2003	89.0 F	Twitter	41.0 F

Figure 2: Out-of-domain performance for several NLP tasks, including POS tagging, phrase structure (PS) parsing, dependency parsing and named entity recognition. Accuracies are reported in percentages; phrase structure parsing and NER are reported in F1 measure.

While many techniques are currently under development in the NLP community for domain adaptation (Blitzer et al., 2006; Chelba and Acero, 2006; Daumé III, 2009; Glorot et al., 2011; Yang and Eisenstein, 2014), including leveraging fortuitous data (Plank, 2016), they often require specialized expertise that can be a bottleneck for researchers in the humanities. The simplest and most empowering solution is often to create *in-domain* data and train NLP methods on it directly; in-domain data can substantially increase performance, almost to levels approaching state-of-the-art on newswire. When adding training data of Early Modern German and adding spelling normalization, Scheible et al. (2011) increase POS tagging accuracy on Early Modern German texts from 69.6% to 91.0%; when Moon and Baldridge (2007) train a POS tagger on Middle English texts, this pushes their accuracy from 56.2% to 93.7%; when Derczynski et al. (2013b) train a POS tagger directly on Twitter data, this increases accuracy from 73.7% to 88.4%. In-domain data is astoundingly helpful for many NLP tasks, from part-of-speech tagging and syntactic parsing to temporal tagging (Strötgen and Gertz, 2012).

The difficulty, of course, is that training data is expensive to create at scale since it relies on human judgments; and the cost of this data scales with the complexity of the task, so that morphosyntactic or semantic annotations (which require a holistic understanding of an entire sentence) are often prohibitive. Few projects achieve this scale for domains in the humanities, but when they do, they have real impact – these include WordHoard, which contains part-of-speech annotations for Shakespeare, Chaucer and Spenser (Mueller, 2015); the Penn and York parsed corpora of historical English (Taylor and Kroch, 2000; Kroch et al., 2004; Taylor et al., 2006); the Perseus Greek and Latin treebanks (Bamman and Crane, 2011), which contain morphosyntactic annotations for classical Greek and Latin works; the Index Thomisticus (Passarotti, 2007), which contains morphosyntactic annotations for the works of Thomas Aquinas; the PROIEL treebank (Haug and Jøhndal, 2008), which contains similar annotations for several translations of the Bible (Greek, Latin, Gothic, Armenian and Church Slavonic); the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria, 2010); the Icelandic Parsed Historical Corpus (Rögnvaldsson et al., 2012), and Twitter, annotated for part-of-speech (Gimpel et al., 2011) and dependency syntax (Kong et al., 2014).

The availability of these annotated corpora means that we have the ability to train NLP tools for some dialects, domains and genres in Ancient Greek, Latin, Early Modern English, historical Portuguese, and a few other languages; this doesn't help the scholar working on John Milton, Virginia Woolf, Miguel Cervantes, or the countless other authors and genres in the long tail of underserved domains that researchers are increasingly finding high-quality NLP useful to help analyze. In this talk, I'll argue for an alternative: an open repository of linguistic annotations that scholars can use to train statistical models for processing natural language in a variety of domains, leveraging information from complementary sources (such as the works of Shakespeare) to perform well on a target domain of interest (such as the works of Christopher Marlowe). What this repository critically relies on is the expertise of the individuals who simultaneously are the consumers of NLP for their long-tail domain and are in the uniquely best position to create linguistic data to support their own work—and in doing so, can help develop an ecosystem that can support the work of others.



Citation	Task	In domain	Accuracy	Out domain	Accuracy
Rayson et al. (2007)	POS	English news	97.0%	Shakespeare	81.9%
Scheible et al. (2011)	POS	German news	97.0%	Early Modern German	69.6%
Moon and Baldridge (2007)	POS	WSJ	97.3%	Middle English	56.2%
Pennacchiotti and Zanzotto (2008)	POS	Italian news	97.0%	Dante	75.0%
Derczynski et al. (2013b)	POS	WSJ	97.3%	Twitter	73.7%
Yang and Eisenstein (2016)	POS	WSJ		Early Modern English	74.3%
Gildea (2001)	PS parsing	WSJ	86.3 F	Brown corpus	80.6 F
Lease and Charniak (2005)	PS parsing	WSJ	89.5 F	GENIA medical texts	76.3 F
Burga et al. (2013)	Dep. parsing	WSJ	88.2%	Patent data	79.6%
Pekar et al. (2014)	Dep. parsing	WSJ	86.9%	Broadcast news	79.4%
				Magazines	77.1%
				Broadcast conversation	73.4%
Derczynski et al. (2013a)	NER	CoNLL 2003	89.0 F	Twitter	41.0 F

Figure 1. Out-of-domain performance for several NLP tasks, including POS tagging, phrase structure (PS) parsing, dependency parsing and named entity recognition. Accuracies are reported in percentages; phrase structure parsing and NER are reported in F1 measure.

Acknowledgements

Many thanks to the anonymous reviewers for helpful feedback. This work is supported by a grant by the Digital Humanities at Berkeley initiative.

Bibliography

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany, August 2016. Association for Computational Linguistics.

Bamman, D., and Crane, G. (2011) The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer, 2011.

Benne, D. C. (2014). Marrying the benefits of print and digital: Algorithmically selecting context for a key word. In *Digital Humanities 2014*, 2014.

Blitzer, J., McDonald, R., and Pereira, F. (2006) Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

Burga, A., Codina, J., Ferraro, G., Saggion, H., and Wanner, L. (2013). The challenge of syntactic dependency parsing adaptation for the patent domain. In *ESSLLI-13 Workshop on Extrinsic Parse Improvement*, 2013.

Chelba, C., and Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20 (4): 382–399, 2006.

Daumé, H. (2009) Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*.

Derczynski, L., Maynard, D., Aswani, N., and Bontcheva, K. (2013a) Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206, 2013b.

Derven, C., Teehan, A., and Keating, J. (2014). Mapping and unmapping Joyce: Geoparsing wandering rocks. In *Digital Humanities 2014*, 2014.

Galves, C., and Faria, P. (2010). Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/corpus/en/index.html .

Gildea, D. (2001) Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202, 2001.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011) Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume*, Portland, OR, June 2011.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., et al. (2009) The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics, 2009.

Haug, D. TT, and Jøhndal, M. (2008) Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1st June 2008*, pages 27–34, 2008.

Houston, N. (2014) Enjambment and the poetic line: Towards a computational poetics. In *Digital Humanities 2014*, 2014.

Jockers, M. (2013) "Secret" recipe for topic modeling themes. http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/, April 2013.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A (2014). A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October 2014. Association for Computational Linguistics.

Kroch, A., Santorini, B., and Delfs, L. (2004). Penn-Helsinki parsed corpus of Early Modern English. *Philadelphia: Department of Linguistics, University of Pennsylvania*, 2004.

Lease, M., and Charniak, E. (2005). Parsing biomedical literature. In *Natural Language Processing–IJCNLP 2005*, pages 58–69. Springer.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19 (2): 313–330, 1993.

Moon, T., and Baldridge, J. (2007). Part-of-speech tagging for Middle English through alignment and projection of parallel diachronic texts. In *EMNLP-CoNLL*, pages 390–399, 2007.

Mueller, M. (2015). Wordhoard. http://wordhoard.northwestern.edu/, Accessed 2015.

Passarotti. M. (2006), Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In Petrilli Raffaella and Femia Diego, editors, *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006*, pages 187–205. Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 2007.

Pekar, V., Yu, J., El-karef, M., and Bohnet, B. (2014)Exploring options for fast domain adaptation of dependency parsers. *SPMRL-SANCL 2014*, page 54.

Pennacchiotti, M., and Zanzotto, F. M. (2008). Natural language processing across time: An empirical investigation on italian. In *Advances in natural language processing*, pages 371–382. Springer.

Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP. In *KONVENZ*, 2016.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Proceedings of Corpus Linguistics (CL2007)*.

Rögnvaldsson, E., Ingason, A. K., Sigursson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *LREC*, pages 1977–1984, 2012.

Scheible, S., Whitt, R. J., Durrell, M., and Bennett, P. (2011). Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 19–23. Association for Computational Linguistics, 2011.

Simon, R., Barker, E. T. E., de Soto, P., and Isaksen, L. (2014). Pelagios 3: Towards the semi- automatic annotation of toponyms in early geospatial documents. In *Digital Humanities 2014*.

Søgaard, A. (2011). Semisupervised condensed nearest neighbor for part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 48–52. Association for Computational Linguistics, 2011.

Strötgen, J., and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Taylor, A., and Kroch, A.S. (2000) The Penn-Helsinki Parsed Corpus of Middle English. *University of Pennsylvania*, 2000.

Taylor, A., Nurmi, A., Warner, A., Pintzuk, S., and Nevalainen, T. (2006). Parsed Corpus of Early English Correspondence. Oxford Text Archive.

Underwood, T., Black, M. L., Auvil, L., and Capitanu, B. (2013). Mapping mutable genres in structurally complex volumes. In *Big Data, 2013 IEEE International Conference on*, pages 95–103. IEEE, 2013.

Wilkens, M. (2013). The geographic imagination of Civil War-era American fiction. *American Literary History*, 25 (4): 803–840, 2013. 10.1093/alh/ajt045.

Yang, Y., and Eisenstein, J. (2014). Fast easy unsupervised domain adaptation with marginalized structured dropout. *Proceedings of the Association for Computational Linguistics (ACL), Baltimore, MD*, 2014.

Yang, Y., and Eisenstein, J. (2016). Part-of-speech tagging for historical English. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1328, San Diego, California, June 2016. Association for Computational Linguistics.

Les capsules vidéo de MemoRekall: la conjugaison des approches intra et inter-documentaire au service de la médiation numérique des arts de la scène

Clarisse Bardiot
clarisse_bardiot@mac.com
University of Valenciennes, Belgium

En 2004, la Tate Modern se dote d'un nouveau département intitulé *time-based media art*, lequel regroupe des œuvres se caractérisant par leur durée. Dans ce glissement de l'espace au temps, de nouveaux « objets » intègrent les collections muséales, parmi eux les performances. Pourtant, objet immatériel, éphémère, non reproductible, la performance est considérée comme impossible à collecter/collectionner. (Bishop, 2014). Seules restent les traces (photographies, captations vidéo, diagrammes, objets...), les reliques, acquises par les musées et parfois élevées au rang d'œuvres d'art. Soit la documentation mais non l'œuvre elle-même dont l'ontologie résiderait dans le *hic et nunc* de la (re)présentation et la coprésence des acteurs et des spectateurs (Phelan, 1993). Comment préserver une œuvre par nature éphémère, réputée attachée au corps de l'artiste, pour l'intégrer dans une collection permanente et être en mesure de la re-présenter au public ? Quels protocoles mettre en œuvre, quelles conditions réunir pour faire en sorte de répéter à l'infini « ce qui ne reste pas » (Van Imschoot, 2005) ? Ces questions sortent du cadre du musée. Elles concernent l'ensemble des arts de la scène, tant la possibilité du répertoire, de la reprise ou de la reconstitution.

Dans ce contexte, la documentation joue un rôle crucial. Sans elle, impossible de re-présenter une œuvre, de la « remonter » (Laurenson & van Saaze, 2014 ; Bénichou, 2015). Les technologies numériques permettent une réappropriation des contenus sémiotiques des documents, et en particulier de la captation vidéo qui jouit d'un statut à la fois privilégié et controversé. La captation vidéo a été avant la capture de mouvement la réponse à l'archivage de ce document qu'est le corps de l'interprète. Provoquant maints débats depuis les années 60, souvent considérée comme une « trahison » de l'œuvre originale (Melzer, 1995a, 1995b), elle s'est aujourd'hui banalisée. La captation vidéo est devenue un document parmi d'autres, au côté des partitions, des carnets de notes, des photographies, des croquis de scénographies, des documents de production, etc. Pour pallier les insuffisances ou les biais de la captation vidéo, deux types de stratégies sont développées : intra-documentaire (annoter la captation) et inter-documentaire (connecter la captation à un corpus documentaire plus vaste).

L'approche intra-documentaire : annoter la captation vidéo

La captation vidéo est rarement autosuffisante et appelle le commentaire, l'analyse, le décryptage. Comme le montre le récent numéro spécial de *Performance Research* intitulé *On An/Notations*, le phénomène est loin d'être marginal et concerne non seulement la pratique des chercheurs mais aussi celle des éditeurs, des artistes et du public (deLahunta, Vincs & Whatley, 2015). L'annotation désigne « toute forme d'ajout visant à enrichir une inscription ou un enregistrement pour attirer l'attention du récepteur sur un passage ou pour compléter le contenu sémiotique par la mise en relation avec d'autres contenus sémiotiques préexistants ou par une contribution originale » (Zacklad, 2007 : 34). Elle est intra-documentaire au sens où, en marge du document – ici la captation vidéo –, elle en est solidaire tout en proposant un enrichissement, un développement, un complément sur un point ou une séquence précise.

Plusieurs initiatives pour créer des applications dédiées aux besoins spécifiques des arts de la scène se sont développées récemment. L'un des axes exploré est la création de logiciels d'annotation vidéo avec de nombreuses approches, témoignant d'un rapport à l'archive et au document loin d'être univoque (Bardiot, 2015a ; 2015b).

L'approche inter-documentaire : connecter la captation vidéo

La captation suscite également une autre approche, qualifiée d'inter-documentaire (Briet, 1951 : 13). Dans le cas de la captation, l'approche inter-documentaire consiste à connecter le document vidéo à un ensemble d'autres documents (textes, images, sons, en ligne ou hors ligne). Ces « autres » documents permettent de combler les lacunes de la captation, d'élargir le cadre de l'image au contexte, au processus de création, à la parole de toute l'équipe artistique, à la réception du public, de préciser des aspects qui ne sont qu'évoqués. La captation ne devient pleinement intelligible qu'au sein d'un écosystème documentaire qui permet d'établir des correspondances, de corréler les informations, de mailler les traces, soit d'en révéler le contenu sémiotique. L'approche inter-documentaire souligne l'interdépendance entre les documents d'un même corpus ou « dossier documentaire ». Ces derniers ne peuvent uniquement se composer d'un ensemble de pièces cohérentes : il faut également établir des relations *entre* les documents.

Les *Dance Capsules* des spectacles de Merce Cunningham et les *Choreographer's Score* d'Anne Teresa De Keersmaeker sont deux exemples récents relevant d'une approche inter-documentaire. Au lecteur de réagencer les documents, de les articuler les uns par rapport aux autres, de trouver son propre chemin au travers d'un corpus hétéroclite et interdépendant, au travers d'un « dossier » dont la captation peut éventuellement être un élément saillant, mais jamais autonome.

Les capsules de Mémorekall

Approches inter-documentaire et intrad-ocumentaire ne sont pas exclusives l'une de l'autre et la frontière est parfois poreuse. Elles peuvent être utilement combinées et jouer simultanément de différents rapports au document, entre approfondissement et connexion. Elles offrent différentes strates de navigation à l'intérieur et à l'extérieur de la captation vidéo et permettent la réappropriation des corpus documentaires non seulement par les artistes et leurs équipes mais aussi par les conservateurs et le public.

Dans le contexte d'un projet plus vaste (Rekall), et avec le souhait de prolonger mes réflexions sur l'ingénierie documentaire liée à la conservation des arts de la scène via l'instrumentation informatique et le développement d'interfaces homme-machine, j'ai conçu en collaboration avec Guillaume Marais, Guillaume Jacquemin et Thierry Coduys une application en ligne, MemoRekall, mise en ligne en septembre 2015. *Open source* et gratuite, MemoRekall permet d'enrichir une captation vidéo en conjuguant les démarches inter et intra-documentaire, l'annotation et la connexion à d'autres ressources documentaires. Au fur et à mesure de la consultation le lecteur découvre annotations et autres documents, qu'il peut ouvrir et consulter à sa guise, pour revenir ensuite au film. Le temps structure l'ordre d'apparition documentaire, même s'il est également possible d'accéder aux différents contenus par leur organisation spatiale dans l'écran et de s'affranchir ainsi de toute chronologie. Le résultat est une « capsule » (nous empruntons le terme à Cunningham) qu'il est possible d'intégrer à n'importe quelle page web. La capsule comprend la captation ainsi que l'ensemble des annotations et des documents liés (ou des liens lorsqu'il s'agit de ressources en ligne). Elle offre une lecture augmentée de la captation au sein d'un environnement multimodal et peut être actualisée à volonté. L'élaboration d'un outil numérique tel que MemoRekall, et la réalisation de capsules à des fins de conservation, ont pour ambition de déduire de l'articulation des documents l'intention artistique de l'auteur de l'œuvre et d'aider ainsi à établir l'authenticité d'une ré-interprétation via la documentarisation.

L'interface comprend quatre espaces principaux : à gauche, une barre de menu, avec des boutons pour ajouter notes, documents ou liens vers des ressources en ligne, ainsi que quelques fonctions de paramétrages (choix d'un auteur et d'un titre, copyright, export en xml, code d'*embed* pour l'intégration dans une page web) ; au centre, la captation vidéo (provenant de Youtube ou Vimeo) et une *timeline* avec les points d'entrée des annotations et des documents connectés ; à droite, une colonne rassemble l'ensemble des documents liés par ordre d'apparition. Les documents peuvent être de n'importe quelle nature : textes, images, sons, vidéos. Ils sont téléchargés par le créateur de la capsule sur le serveur de MemoRekall. Les ressources en ligne sont disponibles sous forme de liens hypertextes et ne sont pas copiées sur le serveur de MemoRekall, dans le respect des droits d'auteur. Documents téléchargés et liens internet peuvent être annotés dans une fenêtre en pop up. Les manipulations s'opèrent directement dans le navigateur internet et sont sauvegardées automatiquement.

MemoRekall propose une lecture multi-documents, temporelle et spatiale, d'une œuvre, à partir de son enregistrement vidéo. Si cette application peut être utilisée dans de nombreux contextes, elle a avant tout été créée pour les besoins spécifiques du *time-based media art* et tout particulièrement des arts de la scène. MemoRekall se situe au carrefour de la valorisation de documents culturels numériques existants, de la création de contenus culturels enrichis et de la mise en place d'espaces critiques et collaboratifs. MemoRekall est à la fois un outil de création, de conservation, de critique d'art et de publication, offrant divers scénarios d'usages : conservation, exposition/publication en ligne, diffusion, médiation numérique, éducation aux médias... Il s'adresse à différents bénéficiaires, privés, collectifs ou publics. En fonction de chaque scénario, le statut et la nature des annotations comme des documents liés sont différents. Nous avons utilisé MemoRekall dans des classes pour familiariser les élèves (du collège à l'université) avec l'analyse de spectacle, dans des théâtres pour la médiation numérique des œuvres auprès de différents publics, avec des artistes pour documenter leur processus de création ou un spectacle achevé.

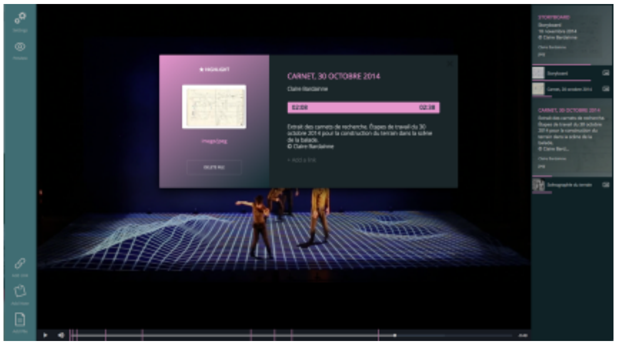


Fig. 1: Capture d'écran présentant l'interface principale de MemoRekall en mode édition.

Spectacle : Mourad Merzouki et Adrien M & Claire B, Pixel, 2014.Capsule réalisée dans le cadre de l'ouvrage numérique La neige n'a pas de sens. Adrien M & Claire B, sous la direction de Clarisse Bardiot, Editions Subjectile, 2016.

Remerciements

MemoRekall a reçu le soutien du Ministère de la Culture et de la Communication (appel à projet services culturels innovants) et de nombreuses institutions culturelles.

www.memorekall.fr

Bibliographie

Bardiot, C. (2015a). « Rekall : an environment for notation / annotation / denotation ». *Performance Research* 20 (6), p. 82-86.

Bardiot, C. (2015b). « Video recording and documentation of scenic arts : from the annotation to the visualization of metadata, the example of the Rekall software », p. 159-68 in *Acoustic Space. Data Drift. Archiving Media and Data Art in the 21st Century* / sous la direction de Rasa Smite, Raitis Smits, et Lev Manovich, 14.

Bénichou, A., éd. (2015). *Recréer / scripter – Mémoires et transmissions des œuvres performatives et chorégraphiques contemporaines*. Les Presses du réel. Dijon.

Bishop, C. (2014). « The perils and possibilities of dance in the museum : Tate, MoMA, and Whitney ». *Dance Research Journal*, 46 (3), p. 63-76.

Briet, S. (1951). *Qu'est-ce que la documentation ?* Édit. Paris.

deLahunta, S., Kim, V., et Whatley, S. (2015). « On An/Notations ». *Performance Research* 20 (6), p. 12.

Laurenson, P., et van Saaze, V. (2014). « Collecting performance-based art : new challenges and shifting perspectives », p. 27-41, in *Performativity in the Gallery. Staging Interactive Encounters* / sous la direction de Outi Remes, Laura MacCulloch et Marika Leino. Bern : Peter Lang Verlag.

Melzer, A. (1995a). « 'Best Betrayal' : the documentation of performance on Video and Film, Part 1 ». *New Theatre Quarterly* 11 (42),p. 147-57.

Melzer, A. (1995b). « 'Best Betrayal' : the documentation of performance on Video and Film, Part 2 ». *New Theatre Quarterly* 11 (43), p. 259-76.

Phelan, P. (1993). *Unmarked : the Politics of Performance*. London ; New York : Routledge.

van Saaze, V. (2013). *Installation Art and the Museum: Presentation and Conservation of Changing Artworks*. Amsterdam University Press.

Salaün, J.-M. (2012). *Vu, lu, su : les architectes de l'information face à l'oligopole du Web.* Cahiers libres. Paris : La Découverte.

Van Imschoot, M. (2005). « Rests in pieces : partitions, notation et trace dans la danse ». *Multitudes* 21 (2), p. 107-16.

Zacklad, M. (2007). « Annotation : attention, association, contribution », p. 29–46 in *Annotations dans les Documents pour l'Action* / sous la direction de Manuel Zacklad et Pascal Salembier. Paris : Lavoisier.

Binary Truths: Developing a Linked Data Model for Historiographical Arguments

M. H. Beals
m.h.beals@lboro.ac.uk
Loughborough University, United Kingdom

Introduction

This paper will argue that historical linked data should not be used solely as means of storing facts but also as a means of improving historiographical discourse by formalising rhetoric, clarifying premises and evidence, and allowing for a distant reading of historical interpretations.

The need for such formalizations can be seen in the following interaction between two historians. An argument is made by one historian with a selection of evidence to support it. Another historian responds with a counter argument and counter evidence. The first becomes frustrated; the second did not understand their argument and their new evidence is irrelevant. The second is equally frustrated as the first has not properly addressed their concerns and instead glosses over them to return to his or her original argument. To the outside observer the conflict appears intractable; neither side is willing to concede the other's points and modify their interpretation of the event or process. This is not a case of academic stubbornness; it is historiographical switchtracking.

Switchtracking (Stone and Heen, 2015) is the result of two similar but non-identical conversations taking place at the same time. In the above, both historians are discussing a single historical problem but have interpreted that historical problem in slightly different ways, responding to each other with own their interpretation of the question in mind. By leaving their specific goals implicit or relying upon ambiguous terminology they have allowed their arguments to be easily misconstrued or misidentified (Godden, 2013). For example, the historical problem "What caused the Salem Witch Trials?" might not only lead to different interpretations—community conflicts, economic disparity, rye-ergot poisoning, religious fanaticism—but also different incarnations of the question itself:

- "What were the causes of the Putnam accusation?" or
- "Why did the Salem Witch Trials begin with the Putnam accusation?" or
- "Why did the Putnam accusation escalate into a wider hysteria?"

The different interpretations that arise from these similar but non-identical questions can lead to historians speaking at cross-purposes and unnecessarily hinder our wider understanding of historical events. The simplest solution is to better define the premises and hypotheses of a given study. However, the challenge of providing a defined, testable hypothesis, combined with the semi-narrative writing style preferred by historians, often precludes this level of clarity. Limited by a perpetually incomplete evidence-base, fuzziness is assumed and allows if not promotes poorly aligned debates.

Linked data may provide a two-fold solution to this problem. Traditionally, clarity and reproducibility in historical research has relied upon three methodological pillars: the quotation, the citation, and the acceptance of interpretive interoperability. The first two work in tandem, providing clear links to or examples of the precise evidence used. Limitations of publishing space have previously reduced the comprehensiveness of these pillars but now the ability to provide targeted hyperlinks and sustainable datasets online allows for a much greater degree of precision. However, the use of page and line numbers, hyperlinks, DOIs, and other edition indicators are inconsistent across the history publications. Likewise, citation standards and allusion conventions differ between publications, subfields, and other communities of practice. It would be difficult, and arguably undesirable, to suggest homogeneity. Without this, however, it is impossible to prevent switchtracking and the misinterpretation as to which precise evidence is being used for which purpose.

The integration of a linked data layer, a meta-document attached to a piece of historical writing, could serve as a remedy to this problem. Bringing together existing ontologies for describing geographical, biographical, and chronological data as well as digital or digitised documents provides a straightforward means for creating strong, definite links between historiography and the data that underpins it. Because such a layer could vary in depth of detail, it could begin with the basic citation information expected of traditional footnoting, but flexibly add layers of detail that would be infeasible in traditional journal or monograph typesetting.

Linked data is already in common usage in certain historiographical and heritage circles, usually in the publication of discrete datasets or in cataloguing digital, digitised and traditional archive collections (Meroño-Peñuela et al., 2013). However, their integration with specific piece of historiographical writing is more complex. At their most basic level, they differ little from traditional citation practises and the added value of precision may not fully compensate for the additional effort in producing this metadata layer. Instead, it is the interpretation of that evidence, and the analytical linkages made by historians, that provide the most significant opportunity for developing historical research. Within the historical and heritage community, linked data is often considered to be limited to 'fact-based' information and cannot convey or represent the analytical frameworks that the narrative provides. Indeed, as of writing, there appears to be only one complete ontology for expressing rhetorical logic, which is poorly maintained with no clear evidence of it being employed in academic debate (Dumontier, 2014). Moreover, it focused upon highly structured logical expression; historical research often relies upon highly fragmented data, requiring vary degrees of informed speculation, which, when undocumented, is the primary cause of switchtracking.

Creating an ontology that can provide definite relationships between evidence, premises, correlations, causations, formal logical deductions and speculative interpretations would allow historians to maintain the semi-narrative writing style expected of historical research but add a layer of unequivocal—if less poetic—statements that provide unambiguous statements of their argument and its components. Beyond serving as a mechanism for researchers to refine their argumentation, presenting a complex historiographical interpretation as a collection of interconnected relationships—literal and rhetorical—would allow for a computational comparison of arguments; similar conclusions could have their evidences combined whereas contradictory interpretations would have a clear set of evidences and premises from which to begin an investigation of divergent views.

This paper will therefore discuss the issues surrounding the creation of an ontology that combines well established ontologies regarding historical evidences and document provenance alongside rhetorical relationships between premises and conclusions. It will demonstrate how one might create a graphical representation of a narrative text and vice versa. The paper will be presented in both semi-narrative long form and RDF triples.

Bibliography

Dumontier, M. (2014). The Semantic Science Integrated Ontology http://semanticscience.org/resource/SIO_000261.rdf (accessed 1 November 2016).

Godden, D. M. (2013) Arguing at Cross-Purposes: Discharging the Dialectical Obligations of the Coalescent Model of Argumentation. *Argumentation*, **17**(2): 219–43 doi:10.1023/A:1024032009784.

Meroño-Peñuela, A., Schlobach, S. and Harmelen, F. van (2013). Semantic Web for the Humanities. In Cimiano, P. (ed), *Proceedings of the 10th Extended Semantic Web Conference, ESWC 2013, Montpellier, France, May 28-30, 2013, Lecture Notes in Computer Science*, vol. 7882. pp. 645–49.

Stone, D. and Heen, S. (2015). *Thanks for the Feedback: The Science and Art of Receiving Feedback Well*. New York: Viking.

The Pushkin Digital Project

Gabriel Belyak
gabriel.belyak@gmail.com
Saint Petersburg State University, Russia

The Russian academic approach towards editing of the nineteenth-century classics, in general, inherits German critical editing tradition. Typically, the edition will include three basic sections: 1) critically prepared definitive text of the work, 2) full critical history with the list of all the discrepancies between its versions 3) commentaries including a description of the manuscript and printed sources of the text, history of its creation, a detailed historical and literary commentary based on all the existing research about it, and finally explanatory notes to the text.

All the works of Dostoyevski, Chekhov, Turgenev and other XIX classics are published this way.O nly one of the 19th century classics still does not have his collected works academically edited. This exception is Pushkin, whose exclusive place in Russian culture can be compared to Shakespeare's in English, Dante's in Italian or Goethe's in German culture. The reasons behind this phenomenon are simply historical. The first attempt to publish a complete critical edition of Pushkin's works was interrupted by the revolution and the civil war, and the second was distorted by the personal order of Stalin. The third attempt is being carried out now, when the government's interests disregard any cultural values. The work towards finally making this edition is conducted at the Institute of Russian Literature of the Russian Academy of Sciences (Pushkin House). Today there are only four volumes published but the amount of copies is 500 in average, and they are marketed almost exclusively in St. Petersburg and Moscow.

The theme of this conference is Access. Low availability of any printed edition today can be easily compensated by publishing it online. However, when we speak about critical editions, the problem of accessibility obtains an entirely different aspect. To use all the information the existing critical edition of Pushkin's works provides, its readers should have a special philological training. and if they have one, in order to actually conduct any work with the materials provided they will require access to Pushkin's autographs, books, magazines of the Pushkin's era and to all researches and papers upon which the commentaries are based. And it goes without mentioning that readers are expected to have a grasp of the main cultural realities of Pushkin's epoch, and to be aware of Pushkin's social and political surroundings. All this makes an anticipated number of readers for this edition only slightly higher than the number of experts who participated in its preparation. Unfortunately, all this can be applied to almost all annotated critical editions. Yet it is obvious that the information they contain can be interesting and useful for every reader. The question is how to make academic publication available for everyone without compromising the scope and depth of its material? And that is the main challenge that stands in front of Pushkin Digital Project.

Developing this project we were thinking not only about how to publish Pushkin's manuscripts, their decipherment and critical descriptions along with the text and the commentary; our ambition was to create an edition that unlike its prototype, which required a certain scientific competence of its reader, would, on the contrary, be able to form such a competence. Our electronic publications were to be addressed not only to researchers and students, but to anyone who would like to have a deeper look into Pushkin's world.

To accomplish that goal, we chose a simple and understandable architecture, consisting of three main operational modes: Text - Commentaries - Manuscripts. Each of these three sections has an autonomous meaning, and could be the sole object for a special digital edition. However, within the framework of our project, it was imperative for all three parts to be closely linked and cross-referenced creating an informational entity, where the text, its history and its context are inseparable from each other.

The Text section of the edition contains the definitive text of the work with two main interacting tools. One tool gives the user the ability to read all the commentary notes upon the text (which may include some external links to music associated with the text and images relevant to this fragment). The other shows all the lines that had any variations along the creation of the text and links to the digitised sources of these variations in section Manuscripts.

The section Manuscripts is a digital representation of autographs and/or printed sources with full transcription and the estimated chronological sequence of author's editings. When working with Pushkin's manuscripts, the presence of a full transcription has a special value due to the extreme intricateness and illegibility of Pushkin's drafts (less than a dozen specialists in the world are able to read it). This section contains a detailed paleogeographic description of each manuscript together with the history of its creation.

The Comment section - is the most voluminous section of the website. Academic historical and literary commentary on Pushkin's works is formed as a reflection and synthesis of all existing studies upon the subject (there is a special department of the Pushkin House dedicated to accounting, describing, and cataloging every book referring to Pushkin or his works (there is no digital version). In the paper edition, this section appears as a coherent and undivided text sometimes up to hundreds of pages. Without knowing its structure and contents in advance, an unprepared reader will easily get lost in it,, and this is if the reader wasn't already scared away by the scientific language in which the text written. Therefore, the digital version of this section was not only rewritten but also splitted into smaller thematic divisions, to allow the reader to navigate easily through all available material and to find quickly what he was looking for. Another problem of the paper edition was a wast amount of proper names, titles and concepts encountered in the commentaries,the meaning of which could be unknown, so that the commentaries themselves needed to be annotated. For the sake of educational functionality, they were to be expanded with an array of background information presented as hyperlinks. That enabled us to engage the full texts of the studies (articles and books), the material of which was reflected in the commentary, as well as those literary texts (including rare editions of XVIII-early XIX century), that the commentaries indicated as significant to the main text. After finally adding audio, different illustrations, and explanatory materials for all the names, facts and places that may be unknown we came from digitising a specific edition of Pushkin's works towards creating an academical encyclopedia for each text included in it.

An encyclopedia of this kind not only extends the perception of the unprepared reader, gradually luring him or her further and further towards the development of the material and teaching a scientific approach towards the text, but at the same time it provides a specialist with all the background and tools required for further studies. Thus, the problem of access to academic publications is solved not in one but in two of its aspects: providing both accessibility and apprehensibility (obtainability) at the same time.

In conclusion, we would like to underline that everything written above is a description of a developing project. At the moment, we have only four dramatical pieces available on our website (the so called Small tragedies). How long would it take to prepare a publication of Pushkin's complete works, we do not now. But in this process, we will certainly face many new problems, which on the larger scale can be described as finding an adequate digital equivalent for different pieces of academical knowledge. To accomplish that task or at least to find a proper vector towards it, it is imperative to become a part of the DH community whose main concerns are in the same field.

Using Methods of Computational Linguistics for Resolving the “Homeric Question”

Christoph Beierle
christoph.beierle@fernuni-hagen.de
University of Hagen, Germany

Norbert Blößner
n.bloessner@fu-berlin.de
Freie Universität Berlin, Germany

Sebastian Kruse
sebi.sk@gmail.com
University of Hagen, Germany

Abstract

The 'Homeric Question' today is no longer a question about Homer as a person, but a question of the genesis and history of the early Greek epic texts. It is still unresolved. Three mutually incompatible Homer theories (Analysis, Neoanalysis and Oral Poetry Theory) compete, which are based only on manual selections of the text, and this fact seems to be part of the problem. In this paper, we report on the development of a toolbox providing methods of computational linguistics intended to improve our capacity to examine texts in their entireity.

Introduction

The Greek epics, traditionally associated with Homer and Hesiod, play a special role among the early texts of Europe. Not only does European literature begin with these poems, but they also mark the transition from oral tradition to written texts. This intermediate position poses very special problems to every philologist, which are best known under the name 'Homeric Question' (going back to Friedrich August Wolf's Prolegomena ad Homerum, 1795): Have these texts been orally composed and later written down (as Oral Poetry Theory claims)? Or are they poems written by a single great poet experienced in oral tradition (as Neoanalysis maintains)? Or do they combine different passages stemming from different times and poets, compiled later (as Analysis assumes)? The core question is: How can we know?

All three theories present evidence for their findings, but this evidence consists of preselected material – obviously selected according to the principle to present what fits best to the own theory. In order to improve our view, it is useful to look at the complete data instead. That is why as early as the 1970s, the University of Regensburg launched a computer aided project aiming at providing the linguistic data needed for an overview. The project was founded by Ernst Heitsch and Xaver Strasser, and its conception and aims are precisely described in Strasser's dissertation thesis (1984).

Oral texts (as we know from Parry, etc.) are constructed not from single words (Lemmata), but from repeated word connections, which the oralists call 'formulae', but a neutral observer would better call 'iterata' (= lat. 'repetitions'). It has been known that repetitions are a key component of testing Homer theories since the 19th century, but at that time there was no reliable way to collect all of the required data. The Regensburg Project created the first complete directory of Epic repetitions (iterata), based on a lemmatized concordance of all Epic word forms. In the "Regensburger Iteratenverzeichnis" (RIV), which is still unpublished, 'iterata' are defined as semantic and syntactic meaningful phrases that occur at least twice in the corpus. It is e.g. possible a) to find passages that are interlinked by the usage of the same iterata throughout the corpus, and b) to have meaningful information about the usage (e.g. frequency, compactness) of words and phrases.

Therefore, the RIV offers the possibility of collecting and presenting all iterates of a certain type, i.e. of specific frequencies or distributions over the epic texts.

The new possibilities have been used for collecting and researching a complete group of iterata: the so-called 'Singulaere Iterata der Ilias', i.e. those repetitions which remain unparalleled in the Iliad. The results have been published in four dissertations (Ramersdorfer, 1981; Csajkas, 2002; Blößner, 1991; Roth, 1989). This group of repetitions is of special interest, because the *Iliad* is the largest and (according to common opinion) oldest of our epic texts. Therefore, an Oral Theory would expect that it is very small, because why should 'old formulae' be so rare in our oldest and largest text? However, this group contains 3,739 Iterata (out of 18,961 Iterata in sum), and in addition, linguistic and semantic research gives evidence in many cases that Oral Theory assumptions do not really explain the facts. It looks as if the claims of the Oral Theory are fundamentally based on (wrong) generalizations of some (correct) results. But also the Neoanalyst position is weakened by demonstrations that, in some hundred cases at least, passages of the *Iliad* presuppose the knowledge of 'younger' texts. These results do not only diminish the weight of widely spread theories, but offer concrete data on which new, and more reliable, theories can be built (cf. Blößner, 2006).

Since these examinations, the methods of Computational Linguistics have improved a lot as has the processing power of today's computer hardware. This paper presents an approach to finding further subsets of iterata that continue and extend the idea of improving Epic theories.

Extending the idea of the 'Singulaere Iterata of the Iliad'

The search aims at finding passages in the text which react to each other. With these results, existing theories can be tested and better ones can be built.

Searches of this kind are applications of the scholars' implicit knowledge. So in order to be able to create a computer assisted system this expert knowledge has to be transferred to describable rules and algorithms. The idea of the 'singulaere Iterata', e.g., defined a "conspicuousness" of a phrase that has an unexpected distribution (contrary to the expectation of some theories). This heuristic was proven valid by the results of the four dissertations mentioned above.

Next we will present some ideas that have a well-known foundation within Computational Linguistics but also can be seen as an extension of the 'singulaere Iterata' idea.

The SIoI compares frequencies between two corpora, the Iliad and the complete corpus but the Iliad, where one frequency is very rare (one). So it can be seen as a subset of the **frequency list comparison**, which searches for terms that differ largely between two (or more) corpora. This method uses the frequency class which for a given corpus *K* and a term *t* is defined as

$$HK_t^K = \left\lfloor 0.5 - log_2 \frac{freq_t^K}{freq_{\alpha^K}^K} \right\rfloor$$

where $freq_t^K$ is the frequency of *t* in *K* and α is the most frequent term in *K*. Now one could search for Iterata *I* where

$$HK_I^{notIliad} \gg HK_I^{Iliad}$$

which obviously extends the SIoI.

With the method of the frequency list comparison one could especially search for iterata that are rare in the *Iliad* but frequent outside of it, but it is hard to decide whether they are of significance regarding the search for parallel passages. So we could use the density of the occurrences outside of the *Iliad* as a filter for this class of Iterata. A well-known metric for this question is the **Chi-squared** test which for partitions *R* (e.g. overlapping passages of 300 verses ignoring book boundaries) of a corpus *K* and an iteratum *I* is according to (Rayson et al., 2004) defined as:

$$\chi_I^2 = \sum_R \frac{\left(freq_I^R - exp_I^R\right)^2}{exp_I^R}$$

A high Chi-squared value now suggests that this iteratum is compact in the sense that many of its occurrences are close to each other in respect to the overall corpus.

Another approach to a density filter is the **log-likelihood-ratio** as proposed in (Dunning, 1993) that according to (Moore, 2004) is also valid for rare events (in this context iterata with low frequency). For partitions *R* of a given corpus *K* and an Iteratum *I* it is defined as:

$$\begin{aligned} sig_I^{R,K} &= 2 \cdot \left[\left(freq_I^R \cdot \log \frac{freq_I^R}{exp_I^R} \right) + \left(freq_I^{K/R} \cdot \log \frac{freq_I^{K/R}}{exp_I^{K/R}} \right) \right] \\ exp_I^R &= freq^R \cdot \frac{freq_I^K}{freq^K} \\ exp_I^{K/R} &= freq^{K/R} \cdot \frac{freq_I^K}{freq^K} \end{aligned}$$

Again the expected occurrences of an iteratum in a given passage are compared to the actual frequency. But also the same metric is applied to the rest of the corpus aside from the considered passage. Again a high log-likelihood-ratio value indicates that the iteratum is compact in this passage. Also it is possible to find passages where an iteratum is more frequent than expected.

An issue may arise as this metric is rather complicated from the perspective of a scholar in the humanities and it remains to be well explained; issues like this have been addressed in the ”ACID for the Humanities" of the DARIAH project (Büchler, 2013) and also in the conclusions of Bestgen, 2013.

First results and conclusions

The ideas presented above have been fully implemented, and this implementation has been used in first applications. It has been suggested to use the Chi-squared test and log-likelihood-ratio to test the validity of an assumption of the Oral poetry, which says that the singer could choose freely from a given set of phrases while performing. This should lead to a rather equal distribution of highly frequent terms. Using our implementation, first results raise concerns regarding the validity of this Oral poetry thesis as the analysis shows that there are many high frequency Iterata that are also “compact” in some parts of the works. Further and more philological work has to be done by analyzing those iterata to find out whether there are semantic reasons for this, and to be able to explain the passages that differ strongly. In general, we expect that the further development of our implementation and its application to different theses proposed by the Homer theories will lead to new insights into the problems named the 'Homeric Question'.

Bibliography

Bestgen, Y. (2014): Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing* 29(2), 164–170.

Blößner, N. (1991): *Die singulären Iterata der Ilias. Bücher 16–20*, Stuttgart (Teubner).

Blößner, N. (2006): Relative Chronologie im frühgriechischen Epos. Eine empirische Methode und erste Ergebnisse, in: *Geschichte und Fiktion in der homerischen Odyssee*, hg. v. A. Luther, München (Beck), 19-46.

Büchler, M. (2013): *Informationstechnische Aspekte des Historical Text Re-use*. PhD thesis, Universität Leipzig.

Csajkas, P. (2002): *Die singulären Iterata der Ilias. Bücher 11--15*, München (Saur).

Dunning, T. (1993): Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1), 61–74.

Janko, R. J. (1982): *Homer, Hesiod and the Hymns*, Cambridge (Cambridge University Press).

Moore, R. C. (2004): On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 333–340. ACL.

Pavese, C. O. / Boschetti, F. (2003): *A Complete Formular Analysis of the Homeric Poems*, Vol. I--III, Lexis' Research Tools.

Ramersdorfer, H. (1981): *Singuläre Iterata der Ilias. Alpha–Kappa*, Königstein/Ts. (Hain).

Rayson, P., Berridge, D., and Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004)*, pages 926–936. Presses universitaires de Louvain (PUL).

Roth, P. (1989): *Singuläre Iterata der Ilias: Phi--Omega*, Frankfurt a.M. (Athenäum).

Strasser, F. X. (1984): *Zu den Iterata der frühgriechischen Epik*, Königstein/Ts. (Hain).

Rewards: Books, Boundedness and Reading in Participatory Culture

Kathi Inman Berens
kathiberens@gmail.com
Portland State University, United States of America

As a professor whose appointment encompasses both electronic literature and printed book publishing, I think a lot about interactive reading and reward systems. Reward systems are fundamental to videogames, where the consequences of reader/gamer choice materialize the stakes of interpretation (Juul, 2005). Computational reader response critics like Bell, Ensslin, Bouchardon and Saemmer frame their discussions of reading as medium-specific and, as I have argued, device specific (Berens). This emergent subfield is "literary interface criticism" (Pold), informed by comparative textual studies (Hayles and Pressman), performance studies (Fletcher), and interface criticism (Galloway; Critical Code Studies Working Group; Chun).

The book publishing world is only now, in the age of mobile-first web access, beginning to reckon with the implications of readers' daily exposure to pervasive, quotidian experiences of interactive reading and gaming environments in video games and social media. These entail nonhuman disruptions in the value chain of traditional industrial book production and distribution. John Maxwell observes that "large swaths of DH [digital humanities] practice overlap or are adjacent to practices in [book] publishing (e.g. markup, database design, user experience design, editing), yet publishing studies and the digital humanities often appear to run at right angles to one another. There is surely an opportunity for complementary work here." My object in this talk is to articulate points of contact between DH and book publishing. Literary interface criticism offers a window onto interactivity and reward systems that extend how we understand the actions of reading: physical, cognitive, social. A book's medium-specific affordances (random-access device; portable; cheap; no Digital Rights Management) are attractive. But how might books take advantage of digital interactivity to provide more rewarding interaction than the "reflowable content" of an e-reader platform?

"Playable books"—the subject of my monograph-in-progress—are part of the "complementary" space Maxwell identifies between book publishing and DH. A playable book is a story object that can be held in human hands, requires physical interaction between human and computer to render, and outputs a story experience that can be "bound" or is otherwise finite. A printed book with interactive elements is playable (see, for example, Tyehimba Jess's *Olio* and Zachary Thomas Dodson's *Bats of the Republic*); a novel displayed on an e-reader is not; an improvised, participatory story in social media is not. (My monograph situates playable books in both literary games and fanfiction archives and databases: such situation is beyond the scope of this short paper–see Note).

In this paper I compare reading's reward structures in one bestseller and a popular, socially dynamic e-literature. I suggest how physically playful digital interactivity could inform mainstream book production and marketing. *Selp-Helf*, a 2015 *New York Times* bestselling book published by Gallery Books (an imprint owned by Simon and Schuster) is a successful YA [young adult] title. It is designed for hands-on, playful interaction; and its author, YouTube sensation Miranda Sings [Colleen Ballinger] sparked such a successful pre-sale campaign that her book débuted at #1 on the *Publisher's Weekly* Nonfiction Hardcover list, and #1 on the *New York Times*' Advice, How-To and Miscellaneous list, where it remained for eleven weeks. But *Selp-Helf* is just one piece of a successful transmedia campaign spanning a Netflix series *Haters Back Off*, a fifty-seven city comedy tour, and several musical albums. Interactivity, in this case, is spread among various media. Book publishing gets a small slice of the pie.

How might mass market book publishing increase its relevance in the contemporary media ecology? I present a short reading of *Ink After Print* (2012), an interactive story machine installed in public spaces such as a rock festival and public libraries, as a way of suggesting what next-generation story interactivity could look like for book publishers who currently funnel their social activity around book into social media campaigns they don't own, because they are hosted in platforms that dictate the terms of interactivity and serve their own agendas. A book like *Selp-Helf* has potential to benefit from reading rewards that are materialized in the book interface itself rather than a paratextual social media campaign.

Rationale & Analysis

Ink After Print is a full-body, playable literary interface. Exhibited at rock festivals, public libraries and train stations (in a French copy of the Danish original), *Ink* brings full-body haptics to the unbound book in ways that resonate with the embodied online social marketing of YA [young adult] titles. Both *Ink* and *Selp-Helf* ask the reader to do real things in the world, and leave traces of those activities in the literary interface, whether it's navigating through the sea of words in "Ink" and printing the results, or posting photos of oneself when meeting the YA author, or dressing up as a character (in this case, teen girls dressing as Miranda Sings.)

In both cases, writer/readers or "w/readers" (Landau, 1999) are having authentic experiences with literary interfaces. "Spreadable media" (Jenkins, Ford and Green, 2013) is a byproduct that empowers Ballinger's fans to use her book as a springboard to articulate their own perspectives on identity and gender. Use of digital skills is the precondition for fan interaction. Jenkins reminds us that *audiences* are individuals, "produced through measurement and surveillance, usually unaware of how the traces they leave can be calibrated by the media industries." *Publics* are collectives that "actively direct attention onto the messages they value" (166). An entire subculture of book fans—often of young adult literature—is using books as totems around which to build worlds made by and through participatory media; *Selp-Helf* is one strong example. *Selp-Helf* and other YA books like it are centerpieces of book-specific media microecologies with particularized rules of conduct, aesthetics, and dynamic interaction. "Playability" focuses through the book, but exceeds the bounded dimensions of the book itself.

As more book marketing focuses on live events captured for and refracted through social media, this paper proposes that book interactivity should do more to engage the actual practice of reading to draw audiences into memorable relationships with the works. Analysis will focus on how the physical aspects of unbound book reading disclose new quantitative and qualitative shifts in mass market book reading practices.

Book publishers, loathe to develop content that they can't expressly monetize, run cheap social media campaigns in platforms they don't own like YouTube, Twitter, Instagram and Snapchat. That's where book publishers should look to literary experimental pieces like *Ink* for how to create "eventness" (pace Bahktin) around distribution and play beyond social media. This would involve investment in digital-first book design and possibly a reading apparatus that could be physically moved location to location. Such techniques could scale, having a few select interactive reading "shows" that are captured in and for the social media audiences. Book publishers have built the expectation among YA readers that social media is their gathering space. Following the example of *Ink After Print*, publishers could offer actual, embodied, interactive reading experiences.

Ink After Print provides a rich context for readers to experiment with their affective experience of boundedness. The mechanics of *Ink* gameplay are sufficiently challenging that readers might feel a sense of reward in assembling a poem using the hand-held "books"; in this sense, the printed receipt is token of achievement. But it is also a highly portable object and a potential gift: to the ephemeral community of others playing *Ink After Print*, where you can share your poems with others who have played; and to the virtual community where Ink "receipts" are stored in the blog. When I curated a media arts show and exhibited *Ink*, I observed readers also folding their receipts into small objects that they then shared with others. The untrackability of what people do with their *Ink* "receipts" stands in stark contrast to the databased traces of participation left by fans of *Selp-Helf*. While *Ink* does output to a blog, its outputs focus on the words themselves, not the user identity. In this sense, *Ink* resists the types of identity quantification that feeds and funds corporate sponsorship of social media platforms.

Note

[1] Working groups of note in this space are: the Games and Literary Theory group founded by Espen Aarseth, and the Critical Code Studies Working Group. Books of note: Astrid Ensslin's Literary Gaming; Timothy Welsh's Mixed Realism: Videogames and the Violence of Fiction; Anastasia Salter, What Is Your Quest? From Adventure Games to Interactive Books; the excellent collec-tion Analyzing Digital Fiction, edited by Alice Bell and Astrid Ensslin. Alice De Kosnik's Rogue Archives, and Amy Earhart's Traces of the Old, Uses of the New discuss the effect of fan archives and tribute sites that (as Earhart shows) are subject to abandonment, decay and obsolescence.

Bibliography

Aarseth, E. J. (1997). *Cybertext: Perspectives on Ergodic Literature*. Baltimore: Johns Hopkins UP.

Berens, K.I. (2015). "Touch and Decay: Porting Tomasula's *TOC* to iOS," in *The Art and Science of Steve Tomasula's New Media Fiction*. New York: Bloomsbury.

Bouchardon, S. (2013). "Figures of Gestural Manipulation in Digital Texts" in *Analyzing Digital Fiction*, ed. Alice Bell, Astrid Ensslin and Hans Kristian Rustad. New York and London: Routledge Press.

Bouchardon, S.. (2016). "Toward a Tension-Based Definition of Digital Literature." *Journal of Creative Writing Studies*, Vol. 2, Issue 1.

Chun, **W.H.K.** (2016). *Updating to Remain the Same: Habitual New Media*. Cambridge: The M.I.T. Press.

Critical Code Studies Working Group. (2014). http://wg14.criticalcodestudies.com/ (Accessed 30 October 2016.)

De Kosnik, A. (2016). *Rogue Archives: Digital Cultural Memory and Media Fandom*. Cambridge: The M.I.T. Press.

Earhart, A. (2015). *Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies*. Ann Arbor: University of Michigan Press.

Ensslin, A. (2014). Literary Gaming. Cambridge: The MIT Press.

Fletcher, J. "Introduction." *Performance Research Journal*. Vol. 18, No.5. Special Issue: *On Writing and Digital Media*. DOI: 10.1080/13528165.2013.867168

Galloway, A. (2012). *The Interface Effect*. New York: John Wiley and Sons.

Juul, J. (2005). *Half-Real: Video Games Between Real Rules and Fictional Worlds*. Cambridge: The M.I.T. Press.

Hayles, N. K. (1999). *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press.

Hayles, N. K. (2004). "Print is Flat, Code is Deep: The Importance of Media-Specific Analysis." *Poetics Today.* DOI: 10.1215/03335372-25-1-67. Accessed 30 October 2016.

Hayles, N.K. and Pressman, J. (2014) *Comparative Textual Media*. Minneapolis: University of Minnesota Press.

Hunicke, R., LeBlanc, M. and Zubek, R. "MDA: A Formal Approach to Game Design and Game Research." http://www.cs.northwestern.edu/~hunicke/MDA.pdf (accessed 5 March 2016).

Jenkins, H., Ford, S., Green, J. (2013). *Spreadable Media: Creating Value and Meaning in a Networked Culture*. New York: New York University Press.

Maxwell, J. "Publishing Education in the 21st Century and the Role of the University." *Journal of Electronic Publishing*, Vol. 17, Issue 2. Spring 2014.

Pold, S. (2016). "Ink After Print: Literary Interface Criticism." *Journal of Electronic Publishing*: Vol. 19, Issue 2: Disrupting the Humanities: Towards Posthumanities. DOI: http://dx.doi.org/10.3998/3336451.0019.207

Pold, S., Anderson, C.U. (2012). *Ink After Print*. http://www.inkafterprint.dk/

Pold, S., Anderson, C.U. (2014). "Post-digital Books and Disruptive Literary Machines: Digital Literature Beyond the Gutenberg and Google Galaxies." *Formules/Revue Des Creations Formelles*.

Pressman, J. (2009) "The Aesthetic of Bookishness in Twenty-First Century Literature." *Michigan Quarterly Review* Vol. XLVIII, Issue 4: "Bookishness: The New Fate of Reading in the Digital Age."

Saemmer, A. (2013). "Hyperfiction as a Medium for Drifting Times: a Close Reading of the German Hyperfiction *Zeit für die Bombe*" in *Analyzing Digital Fiction*, ed. Alice Bell, Astrid Ensslin and Hans Kristian Rustad. New York and London: Routledge Press.

Salter, A. (2014). *What Is Your Quest? From Adventure Games to Interactive Books*. Des Moines: University of Iowa Press.

Sings, M. (2015). *Selp-Helf.* New York: Gallery Books.

Tender Claws [Samantha Gorman and Danny Cannizarro]. (2014, 2016). *Pry*. Self-published; distributed through Apple App Store.

Tosca, S. (2013). "*Amnesia*: The Dark Descent" in *Analyzing Digital Fiction*, ed. Alice Bell, Astrid Ensslin and Hans Kristian Rustad. New York and London: Routledge Press.

Welsh, T. (2016). *Mixed Realism: Videogames and the Violence of Fiction.* Minneapolis: University of Minnesota Press.

Object Classification in Images of Neoclassical Artifacts Using Deep Learning

Bernhard Bermeitinger
bernhard.bermeitinger@uni-passau.de
Universität Passau, Germany

Simon Donig
simon.donig@uni-passau.de
Universität Passau, Germany

Maria Christoforaki
maria.christoforaki@uni-passau.de
Universität Passau, Germany

André Freitas
andre.freitas@uni-passau.de
Universität Passau, Germany

Siegfried Handschuh
siegfried.handschuh@uni-passau.de
Universität Passau, Germany

Classifying aesthetic forms – a methodology at the heart of art history

The transformation of aesthetic styles has been at the heart of art history since its inception as a scholarly discipline in the late eighteenth century. Analyzing the single artifact and the carefully curated corpus have been the techniques for crafting hermeneutic understanding for such processes of change. Recently new instruments based on statistical techniques empower us for a fresh take on bodies of sources once disregarded as second tier complementary sources such as for instance very large corpora.

The Neoclassica research framework

The *Neoclassica* research framework (Donig et al., 2016) was conceived to provide scholars with new instruments and methods for analyzing and classifying artifacts and aesthetic forms from the era of Classicism (ca. 1760–1860). The neoclassic movement was of almost global scale—affecting architecture and design from Sidney to New York, and from Athens to the outreach of the Russian Urals—while relating to a common reference in classical antiquity, therefore making it an ideal topic for studying processes of stylistic transformation.

It accommodates both traditional knowledge representation as a formal ontology and data-driven knowledge discovery, where cultural patterns will be identified by means of algorithms in statistical analysis and machine learning, having in particular the potential to uncover hitherto unknown patterns in the source data. The outcomes of the top-down and the bottom-up approach will be united in a consistent, unified formal knowledge representation.

Motivated by the need to combine object classification with domain knowledge representation, the ontology focuses at the moment on artifacts (in particular furniture and architecture) and their components. Following the preliminary hypotheses that aesthetic forms in furniture and architecture are in closest communication with each other due to constructional commonalities and their shared reference of the Classic, we decided to start developing the knowledge discovery module of *Neoclassica* by classifying artifacts in digital images.

Knowledge discovery

In this paper, we report on our efforts for using deep learning for classifying artifacts in digital visuals. We chose a deep learning approach for our classification method because of its current superiority over other methods and still rising accuracy over the last years in nearly all image classification and object detection challenges.

Initially, we compiled a body of images both from commercial sources such as auction houses, antique dealers and other public sources. Due to the complex copyright situation, this corpus can not be redistributed. In order to make our experiment reproducible and since the *Metropolitan Museum of Art* (MET) has released 375,000 images in the public domain (The Metropolitan Museum of Art, 2017). we assembled a corpus of 379 artifacts relevant to our research. We processed this corpus with the same algorithm as the original proprietary corpus and released the data together with the source code (The Neoclassica Project, 2017).

Classifier description

The main classifier for our experiments is a *Convolutional Neural Network* (CNN). It classifies an input image as a whole.

In a first step, we applied a standard implementation of a CNN (namely *VGG19* (Russakovsky et al., 2015)). The results were not satisfactory for our needs. It classified the type of the object depicted in the image with an accuracy of 0.37.

In a second step, we opted to employ pre-training, a common technique for improving accuracy in neural networks. We experienced that available pre-trained classifiers for generic image classification proved ineffective in our case. Most of them are trained on a specific subset of *ImageNet* (Deng et al., 2009), containing 1000 classes. These classes are broadly spread around everyday objects like dogs and planes. This led us to assume that the amount of very different classes that don't occur in our corpora interfere with the classification. Following that hypothesis, we decided to train the algorithm on a specific subset compiled from *ImageNet* mainly containing different furniture objects like tables, chairs, and cabinets. They sum up to 35,000 images. The first training step with these images resulted in an accuracy of 0.54 of classifying the object correctly.

First layout

The first corpus contained 2,129 images representing 300 European period artifacts mostly in a colored format of highly diverging quality and resolution. They depict the objects fully, partially or are close-up shots of specific forms. We coarsely annotated these images by manually labeling them on the level of folders. The concepts applied during this labeling process are directly taken from the *Neoclassica* ontology and describe concepts for types of artifacts. These concepts were derived from period sources.

The depth of the class hierarchy was partly reflected by the folder structure. The folder labeled "Chest of drawers" contains all instances of this class. Their labels in turn reflect the names of all the sub-classes in the most extensive specification (e.g. semainier, Wellington chest, commode scriban).

After pre-training, the next step was fine-tuning with this corpus. The accuracy was 0.44, the F1 measure 0.44.

Second layout

The second corpus was assembled from open data released by the MET. It contains 1,246 images representing 379 European and American period artifacts ranging roughly from 1780–1840 including some transition pieces, drawings, and prints. They also depict the objects fully, partially, or are close-up shots of specific forms. We used the titles provided by the MET and manually aligned them with the *Neoclassica* ontology.

The overall mean accuracy over all classes was 0.36, the F1 measure 0.21. For the computation of these numbers, all results that are non-computable (due to only having one image in either the train or test set) were removed. These low numbers result from the existence of two many artifacts represented by only one image, thus making a split in training and testing data meaningless. However, applying pre-training using same *ImageNet* corpus as in the first layout yielded a mean accuracy over all classes of 0.59 and a F1 measure of 0.58.

In order to achieve better results and since the classifier classifies the image as whole, we excluded all images that did not depict the whole artifact. We kept multiple copies of the same image if they were used to describe a different but similar object. We split the images depicting multiple objects so that the resulting images represent only one artifact. We also processed these images so that neighboring objects were covered with solid colors. The images that could not be split (e.g. room interiors) were excluded from the corpus.

Using the same settings with the curated corpus and with pre-training we achieved an accuracy of 0.77 and an F1 measure of 0.76.

Challenges

While pre-training and improving the curation process helped us to raise the accuracy, we assume that there is room for improvement.

Parameters to be taken into consideration include the small size of corpora and how to overcome this limitation since this limits the effectiveness of a neural net. Additionally, since pre-training has been proven to enhance the results, it is rational to assume that a pre-training corpus better suited to period artifacts would improve the results further. Third, our experiment was affected by the limitation of the standard implementation of the CNN which classifies the image as a whole and not parts of it.

Outlining parts inside an image and classifying them is a difficult task for machine learning methods. Recently, a new type of neural net emerged that tackles this challenge: *Regional CNN* (RCNN). It is implemented most prominently in an algorithm called *MultiPath Network* (Zagoruyko et al., 2016).

Current improvements: Using a Regional CNN

We are manually annotating regions within the images with classes from the ontology for training a RCNN that locates objects.

The implementation of the RCNN is divided into two steps. First, it detects objects in the image and draws their outline as a polygon. The second step is classifying the outlined objects using the included standard CNN.

Preliminary results of *MultiPath Network* with default pre-trained settings show that the first step of the RCNN already outlines objects in our corpora within reasonable limits. The corpus for pre-training is *COCO* (Lin et al., 2014). Naturally, specific domain objects are not located and the class names are too generic. For our purpose, fine-tuning on a custom annotated corpus is essential. An RCNN requires a more detailed corpus. The exhaustive task of manually draw the objects' outlines within an image promises higher quality in locating objects (first step) and is necessary to classify the objects according to the ontology (second step).

Future work

We decided to take two steps in the near future for improving our results.

First, we are compiling a new corpus to train the RCNN with, avoiding pitfalls like inconsistent quality, heterogeneous image rights and an inadequate distribution of image per class. Here we would like to go a dual approach. Together with domain experts, we intend to collate a corpus from the large repository of a major auction house, providing us not only with a selection of artifacts' images but also with texts to be used in multimodal analysis.

On the other hand, this kind of artifacts may exhibit provenance issues (e.g. heterogeneity or lack of provenance). We will thus compensate for such issues by digitizing a major corpus of neoclassical artifacts forming an ensemble and comprising artifacts in multiple modes having evolved in close reference to each other. Therefore, we have entered a partnership with the Dessau-Wörlitz UNESCO world-heritage site, an almost untouched complex of manor houses and their furnishings in early neoclassical style.

Regarding the annotations, we are developing our own semantic annotation and ontology population tool since January 2017. The tool will create an annotated corpus. The actual annotation process will be conducted in cooperation with emerging domain experts from the chair of Visual Culture and Art History at the University Passau.

Bibliography

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*.

Donig, S., Christoforaki, M. and Handschuh, S. (2016). Neoclassica - A Multilingual Domain Ontology. In Bozic, Mendel-Gleason, Debruyne and O'Sullivan (eds), *2nd IFIP International Workshop on Computational History and Data-Driven Humanities*.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693 LNCS. pp. 740–755 doi:10.1007/978-3-319-10602-1_48.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, **115**(3): 211–252 doi:10.1007/s11263-015-0816-y.

The Metropolitan Museum of Art (2017). The Met Makes Its Images of Public-Domain Artworks Freely Available through New Open Access Policy http://www.metmuseum.org/press/news/2017/open-access (accessed 1 March 2017).

The Neoclassica Project (2017). Neoclassica – A Framework for Research in Neoclassicism http://www.neoclassica.network/resources (accessed 1 April 2017).

Zagoruyko, S., Lerer, A., Lin, T.-Y., Pinheiro, P. O., Gross, S., Chintala, S. and Dollár, P. (2016). A MultiPath Network for Object Detection. *BMVC* http://arxiv.org/abs/1604.02135.

A Database of Online Book Response and the Nature of the Literary Thriller

Peter Boot
peter.boot@huygens.knaw.nl
Huygens ING, The Netherlands

Introduction

The study of literature has traditionally focused on the literary work, and sometimes its author, rather than on the response that works evoked in their readers. The arrival of the computer in the study of literature has not really changed that – perhaps unsurprisingly, as reader response has never been systematically recorded. The fact that readers have begun to document their reading and reading response on websites is therefore very fortunate (Gruzd and Rehberg Sedo, 2012, Maryl, 2008: 390-406). Booksellers' sites such as Amazon and review sites such as Goodreads, as well as weblogs, forums and general-purpose social media sites provide access to first-hand reading reports. Though most research on these sites focusses on (behavior of) users (Nakamura, 2013: 238-243, Thomas and Round, 2016: 239-253), we are beginning to see them being used in literary research (Finn, 2011) .

This paper presents the Online Dutch Book Response (ODBR) database, that was designed to facilitate research into book response. At present, the database holds reviews and other response items from an online bookseller as well as from four Dutch-language mass review sites, including, where available, the information about books, reviewers and sites necessary to put the reviews into context.

To show one type of research that the database supports, the paper displays a clustering of reviews by genre, based on frequently used words. I discuss the clustering and what it suggests for further research.

Content: mass review sites, reviews from booksellers' sites, weblogs

While the ODBR database was designed to hold any kind of (online) book discussion, at the moment it holds 280,000 response items from four mass review sites: three from the Netherlands (hebban.nl, dizzie.nl and watleesjij.nu) and one from Flanders (lezerstippenlezers.be). It also holds 313,000 items from the main online bookseller in the Netherlands, bol.com, and 36,000 brief expert reviews. We are currently working on downloading about 400 book blogs.

Mass review sites are sites where the main focus is on book reviews uploaded by the sites' users. The best known example is Goodreads (goodreads.com) (Thelwall and Kousha, 2016: 972-983). Hebben.nl, for example, is currently the largest mass review site in the Dutch-language area. It is in many respects like Goodreads: users post reviews, they can follow other users and they can create book lists. Other users can respond to the reviews or vote them up or down. There are moderated reading clubs on specific books. Apart from the user-contributed content, Hebban also holds a fair amount of editorial material, among other things expert reviews, blog posts, interviews and giveaways.

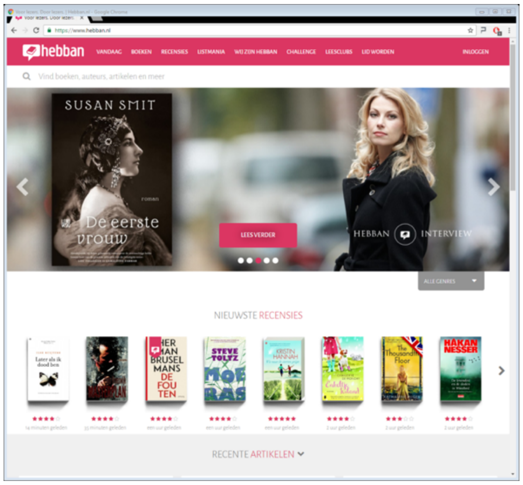


Figure 1. Front page of Hebban.nl

The other three sites are largely similar. Lezerstippenlezers.be ('readers tip off readers') is somewhat simpler in that it lacks the social functions of the other sites. Noteworthy about dizzie.nl is among other things that the site was downloaded for inclusion into the ODBR database only days before it was closed down. The site administrators explicitly stated in their final announcements that no archive would be kept. The ODBR database may be the most complete record of the site's existence. Since then, watleesjij.nu has also been closed down.

Database design and content

The purpose of the database is to facilitate research into online book response. It should be able to store the response texts as well as information about the response's context. Response items are not just reviews. They include book lists, expert reviews, blog posts and review responses and other response types. Ratings and tags are also stored in the database. Figure 2 shows the main entities in the database.

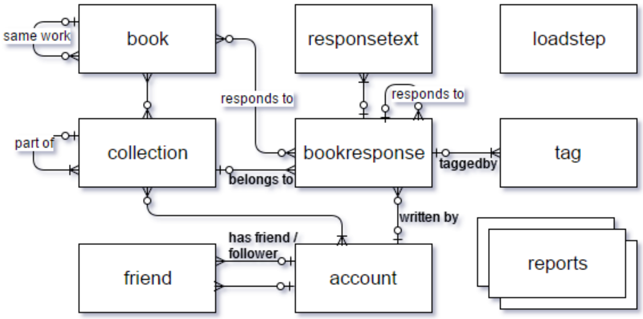


Figure 2. Main entities in ODBR database

The database now contains the following numbers of records: 146,800 books, 58,000 user accounts, 40,000 friendships/followers, and 628,800 book responses. The types of the responses are given in Table 1.

Article type	Number
article	4584
articleresponse	7445
blog	6122
blogresponse	2536
bookdesc	149002
expertreview	36634
list	10280
listresponse	2380
PM	14021
quote	3389
review	382998

Table 1. Article types in ODBR database. PM: private message; Bookdesc: book descriptions provided by sites. The 'response' items (blogresponse, etc.) are responses to a response: a blogresponse is a response to a blogpost.

This large collection of book response items and context information creates many different possibilities for research. First of all, the response texts facilitate the investigation of response to individual books, authors and genres. The texts show the norms that readers apply as well as the way reading affects them. The availability of expert reviews alongside general readers' responses creates the possibility to study differences between (semi-)professional and lay reading. Other types of content enable different lines of research. Information about friends and followers can help investigate the influence of the social environment in reading choices and book appreciation. As many writers use the book review platforms to get in touch with their (potential) readers, the collected data can also provide insight into their marketing strategies. Information about the book lists that readers create (for instance 'to read', 'read in 2014') shows which books are perceived as similar, prompting the question whether they will also be rated similarly. The integration of the discussions and the context information from multiple sites in a single environment that facilitates integrated querying is something that, as far as I know, has not been done before.

Unfortunately, because of copyright and privacy concerns, the database is not accessible over the web. Researchers who are interested in accessing its content are asked to contact the author.

Clustering by genre

Of the research possibilities that the ODBR database offers, here I discuss just one example, an investigation in the word use in reviews by genre. This is an interesting subject, as word usage can be considered as an indication of how people respond to books. I will show a clustering of the genres by word use, which should give a first indication of which genres are perceived by readers as similar.

Dutch publishers use a shared system for classifying their books. This so-called NUR code (an abbreviation meaning Dutch-language Uniform Categorization) covers aspects of format as well as genre. On some of the downloaded sites, books were assigned NUR codes. As the load process of the database tries to merge the book information from multiple sites, NUR codes are available for 75% of the reviews. In the computations, reviews were merged by NUR code. Relative frequencies were computed for all words, these frequencies were then transformed into z-scores. The Euclidean distances between the frequency vectors for the 200 most frequent words were computed and formed the basis for the clustering dendrogram in Figure 3. The figure only shows NUR codes for which there are more than 500 reviews available. For other choices for most frequent words and distance measure, the clustering is largely similar.

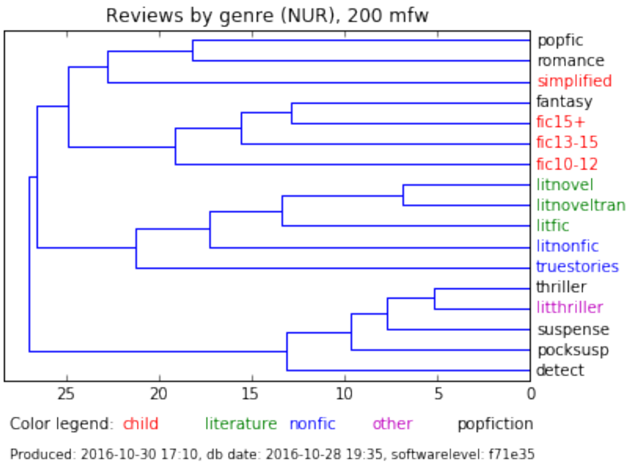


Figure 3. Dendrogram of NUR codes, based on distances between word usage in the corresponding reviews. The colors of the NUR labels reflect a higher-order grouping of genres (see legend)

A number of interesting groupings appear: the literary novel, original and in translation, groups nicely with other literary fiction. At a higher scale literary fiction groups with literary nonfiction and, interestingly, with true stories. Fantasy groups with youth literature, maybe a reflection of its popularity among younger readers. A group of general popular fiction and romance is also close to youth literature. Perhaps the most remarkable in the clustering is the location of the literary thriller. It sits squarely on one branch with the other suspense books such as the regular thriller and the detective. Those who doubt whether the 'literary' in the literary thriller is more than a marketing label, will see their views confirmed by this clustering.

Looking for an explanation, we can drill down to look at the individual words underlying the dendrogram and see for example how readers of (literary) thrillers talk about plot and plot lines, as one would expect. They also speak about 'main characters'(plural), while people who discuss literature use the singular 'main character'. That suggests an interesting difference between literature on the one hand and (literary) thrillers on the other. To praise a book, readers of literature use 'beautiful' and 'nice', readers of (literary) thrillers use 'good' or 'great'. For other observations I am still looking for an explanation. Why, for example, do people who discuss literature use more personal pronouns? That question, like many others, requires further investigation.

Conclusion

Until now, most humanities-oriented researchers that have worked on online book discussion sites and communities have taken a qualitative approach (Fister, 2005: 303-309, Foasberg, 2012). The ODBR database is meant to facilitate quantitative research into online book discussion and, through the lens of online book discussion, into literature, both with respect to its effects on readers and as a social phenomenon. The rich data model and the large quantity of available data should provide support for both language and network oriented research approaches.

Bibliography

Finn, E.F. (2011). The Social Lives of Books: Literary Networks in Contemporary American Fiction. PhD, Stanford University.

Fister, B. (2005). "Reading as a contact sport". Reference & User Services Quarterly, 44 (4): 303-09.

Foasberg, N.M. (2012). "Online Reading Communities: From Book Clubs to Book Blogs". The Journal of Social Media in Society, 1 (1).

Gruzd, A. and Rehberg Sedo, D.N. (2012). "# 1b1t: Investigating Reading Practices at the Turn of the Twenty-first Century". Mémoires du livre, 3 (2).

Maryl, M. (2008). "Virtual Communities – Real Readers: New Data in Empirical Studies of Literature" in Auracher, J. and Van Peer, W. (eds.), Virtual Communities – Real Readers: New Data in Empirical Studies of Literature. Cambridge: Cambridge Scholars Publishing, pp. 390-06.

Nakamura, L. (2013). "Words with Friends: Socially Networked Reading on Goodreads". PMLA, 128 (1): 238-243.

Thelwall, M. and K. Kousha (2016). "Goodreads: A social network site for book readers". Journal of the Association for Information Science and Technology, 68 (4): 972-83.

Thomas, B. and J. Round (2016). "Moderating readers and reading online". Language and Literature, 25 (3): 239-53.

Libraries and Digital Research: Sharing the Incubator

Zoe Borovsky
zoe@library.ucla.edu
UC Los Angeles, United States of America

Claudia Horning
chorning@library.ucla.edu
UC Los Angeles, United States of America

Dawn Childress
dchildress@library.ucla.edu
UC Los Angeles, United States of America

As humanities scholars embrace digital research and more content is digitized, tools and methods such as text mining and spatial analysis become increasingly more vital, especially to graduate students who are at a formative stage in their careers. Librarians and library staff are keen to support these researchers and meet the growing demand. To address this demand, UCLA librarians developed DResSUP (Digital Research Start Up Partnerships), a six-week summer program designed to create partnerships between library staff and researchers. At our institution, a core group of library staff have the expertise to engage with digital research projects beyond the initial phase of data discovery. Therefore DResSUP was initially built around this core group, allowing us to build a local community of practice and test our capacity for supporting digital projects. We have purposely kept the program small; working with a cohort of four to six graduate student "partners" each summer. Our strategy was, no doubt, also influenced by discussions of sustainability in the Digital Humanities community (see, for example, Maron and Pickle, 2014). What if we built a new program as though we were already in the sunset period? How could we minimize costs, maximize impact, build deep connections with researchers, respect ethical issues of graduate student labor, and facilitate engagement between library staff and researchers? We knew about minimal computing, but could we create a "minimal program" and still have an impact?

The goal of the program is to provide graduate students with skills and methods to continue the projects on their own. Rather than doing projects for them, we focus instead on teaching students to start with a small sample of their own dataset, working it through the research data lifecycle: collecting, cleaning, and analyzing/visualizing data, supplemented by a workshop on project management and the specific tools that are applicable to the students' research projects. In the second half of the program, we focus on the execution of their individual projects, building prototypes and adjusting workflows that allow them to complete their projects independently.

After successfully offering this program for the past two years, we are expanding the program in ways that we hope will increase the library's capacity to support digital scholarship, which will make the program more sustainable over time and allow us to reach more researchers. Perhaps because we took a low-profile, grassroots approach, and because of general resistance to "reskilling" or "retooling" efforts, the most surprising response to DResSUP for us has been the enthusiasm and curiosity of our library colleagues. As they have learned more about the program, they have expressed interest in participating. However, at the same time, they raised legitimate concerns about their ability to support advanced digital research methods. Recognizing that they face a steep learning curve, one that will require additional resources to surmount, we have designed an expansion to the program, developing a second track for professional development for librarians.

Similar to Columbia University Library's Developing Librarian and Indiana's Research Now librarian training programs, we will take a project-based approach to professional development for librarians. But, rather than developing a separate program, librarians will participate in the extended DResSUP program, starting earlier in the year, gradually merging with the summer DResSUP graduate students. We will begin working with a small cohort of librarians in January 2017, who will follow the same curriculum as the graduate students. By Summer, the librarian cohort will have a plan for a collaborative digital research project.

Our short paper will begin by presenting DResSUP as it has functioned the past two years and our plans for the next three years, based on preliminary findings from this second librarian track of DResSUP which, by DH2017, will be well underway. We reason that by building capacity in our library community among permanent staff we can extend and grow DResSUP in ways that are beneficial to both researchers and those responsible for building and maintaining the infrastructure that supports those endeavors. We will then discuss results from an open-ended questionnaire that we have designed to surface some of the issues around collaboration between librarians and digital humanities researchers (Siemens et al, 2011) and provide concrete suggestions for faculty, library administrators, and library staff members. Typically, the library's role in a digital project is negotiated when faculty are writing grant proposals and seek assistance with a required data management plan. We will argue that this strategy puts librarians at a disadvantage, having to function as gatekeepers in a bargaining process when researchers are hard pressed to stretch precious resources between their research goals and an institution's needs to cover the costs of supporting research functions.

Finally, we will discuss the ways in which merging professional development for librarians with graduate students provides advantages by focusing library engagement efforts at an earlier stage, i.e., when graduate students begin research for their dissertations or when they begin working as research assistants. We will argue that graduate students function as vectors for spreading the message that librarians can provide valuable expertise: they teach undergraduate students, their peers, and their faculty. In this way, librarians can expand their impact, by focusing strategically on graduate students.

Bibliography

Maron, N. L., and Pickle, S. (2014) “Sustaining the Digital Humanities: Host Institution Support beyond the Start-Up Phase.” New York, NY, USA: Ithaka S + R, June 18, 2014. http://digital.library.unt.edu/ark:/67531/metadc463533/m2/1/high_res_d/SR_Supporting_Digital_Humanities_20140618f.pdf .

Siemens, L., et al. (2011) "A tale of two cities: Implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities." *Literary and linguistic computing* 26.3: 335-348.

I Built an App to Revitalize a Language: Now What?

Nicolle Bourget
nbourget299@telus.net
Royal Roads University, Canada

Summary

Indigenous communities are using technology to document languages and support language maintenance and revitalization activities. My research examined how technology has been incorporated into Upriver Halq’eméylem language programs. Participants identified that ICT is being used successfully as a supplementary tool in coordination with specific learning strategies and activities such as story-telling, games, and looking up a word or concept but that the ICT is not being used outside of those specific learning activities. The study indicates that ICT can be a valuable tool in the effort to revitalize a language; however, technology to revitalize languages needs to be carefully planned with post-implementation activities and oversight to ensure that the language continues to grow.

A Real-World Problem

Linguists estimate that 50-90 percent of the 6,000 to 7,000 known languages in the world will disappear in the 21st century (Grenoble, 2011) with Harrison (2007) estimating that loss is occurring at a rate of one language every ten days. Endangered languages are often spoken primarily, or only, by Elders and as fluent Elders are lost, so is the language. Indigenous languages in Canada are not exempt from language shift; indeed, only 50 of the more than 60 Indigenous languages known to have been spoken in Canada exist today and most of these languages are classified as either endangered or already extinct (Kirkness, 1998). Only three (Norris, 1998) or four (Kirkness, 1998) of these languages are expected to survive.

The loss of an Indigenous language is associated with the loss of Indigenous knowledge and culture. Such knowledge systems incorporate social and historical dimensions including social relationships, cosmology or world views, oral history, place names, spiritual relationships, ecological knowledge, oral literatures, and philosophies (Battiste, 2008; Berkes, 1993; Hinton, 2008a, 2008b; Kipp, 2009). These knowledge systems are embedded within the language and the loss of the language results in the loss of the knowledge systems. Language is important to the health of the community and language revitalization has been identified as playing “a vital role in community growth, healing, education, development, strong families and reconnection to the past” (First Peoples’ Heritage, Language and Culture Council, 2010: 7). Language loss and revitalization is a global, real world problem.

Challenges

Challenges to language revitalization include lack of ideological clarification (Dauenhauer and Dauenhauer, 1998); disagreement as to recording or sharing language (Adley-SantaMaria, 1997); differences in personal beliefs (Kroskrity, 2009); economic impacts (Adegbija, 2008; Hornberger and King, 2008; Kroskrity, 2009); the perceived status of a language and the self-esteem of speakers; and feeling shame and embarrassment about the language and culture (Dauenhauer and Dauenhauer, 1998).

The digital divide is another potential barrier to language revitalization. The digital divide separates individuals and communities who have access to technology and those individuals and communities that do not. Exacerbating the issue of the digital divide is that language programs may inadvertently become technology projects which “often focus on providing hardware and software and pay insufficient attention to the human and social systems that must also change for technology to make a difference” (Warschauer, 2004: 6). Discussions around the digital divide must include the technical aspects such as access to technology and problems with infrastructure as well as the social aspects including education in the technology, gender, age, language, economics, and literacy (Warschauer, 2004).

Technology

Information and communication technology has been used with languages since the late 1800s, when audio recordings of Indigenous peoples were made on wax cylinders (Makagon and Neumann, 2008). These recordings allowed for unidirectional activity; that is, individuals could listen to the recording but could not interact with it. Today, the advances in ICT can provide an interactive, bi-directional experience in which users can interact either with the technology or other users. Language can also be captured in context with cultural activities allowing for a deeper understanding of the language. Multimedia applications are becoming increasingly easy to create and allow for the integration of video, audio, pictures, and text, as well as interaction with human beings. Access to databases and dictionaries provides teachers, administrators, and learners immediate access to language at the word, grammatical, and contextual levels. However, Tyler (2002) notes that:

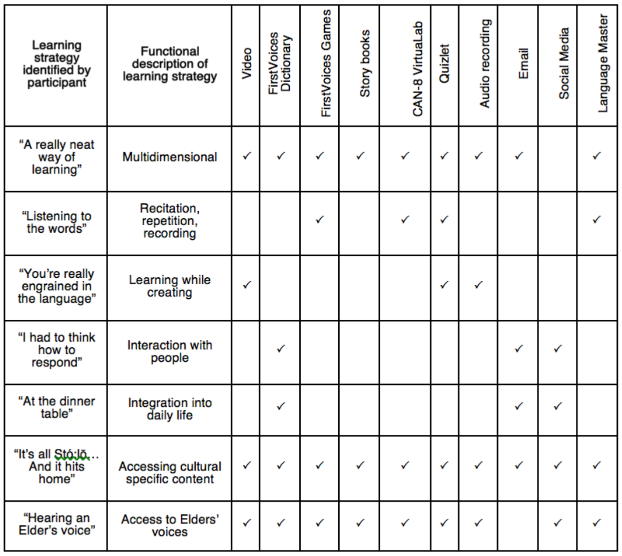
> *"The Internet provides people with a technology that allows them to engage in activities that they have already had ways to engage in but provides them with some added efficiencies and opportunities to tailor their interactions to better meet their needs. However, there is nothing fundamentally different about the Internet that transforms basic psychological or social life.* " (204)

If we apply this statement to Indigenous language programs, technology will be most successful where the language is already being used and where the language is not being used, technology will not increase the usage as is the case of the Upriver Halq'eméylem language community. An endangered language requires very different strategies than a thriving language, and these strategies should determine how digital technologies are used. For example, a community with a thriving language may use technology to encourage conversations between geographically dispersed individuals, to increase the use of language through written communication using email or chat functions, or to provide exposure to the language by posting information on social media sites or blogs. A community with an endangered language may choose to use technology for documentation and archiving so that the language is not lost forever.

My research attempted to understand the effectiveness of technology within an established language program with the goal of providing additional information to help communities that are either considering a language program or have one in flight that uses technology.

Findings

The Upriver Halq'eméylem language community began to use ICT in the mid-1900s to document their language. Over time, the community continued to incorporate ICT and today ICT is an integral tool in the teaching of the language. Table 1 identifies the ten technologies identified as being used with the Halq'eméylem language along the top row and the learning strategies used in the first column.



Learning strategy identified by participant	Functional description of learning strategy	Video	FirstVoices Dictionary	FirstVoices Games	Story books	CAN-8 VirtualLab	Quizlet	Audio recording	Email	Social Media	Language Master
"A really neat way of learning"	Multidimensional	✓	✓	✓	✓	✓	✓	✓	✓		✓
"Listening to the words"	Recitation, repetition, recording			✓		✓	✓				✓
"You're really engrained in the language"	Learning while creating	✓					✓	✓			
"I had to think how to respond"	Interaction with people		✓						✓	✓	
"At the dinner table"	Integration into daily life		✓						✓	✓	
"It's all Stó:lō... And it hits home"	Accessing cultural specific content	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
"Hearing an Elder's voice"	Access to Elders' voices	✓	✓	✓	✓	✓	✓	✓		✓	✓

Table 1: Intersection between ICT and learning strategies as identified by participants

Understanding the how the learning strategies intersected with the technology used provided key information as to how the technology supports language learning. Participants identified that ICT is being used successfully as a supplementary tool in coordination with specific learning strategies and activities such as story-telling, games, and looking up a word or concept but that ICT is not being used to support Halq'eméylem learning activities outside of those specific learning activities. Additionally, participants indicated that ICT that enables human to human interaction has significant potential to contribute to developing fluency but only if the language is already being used.

Table 1 highlights that ICT is rarely used with the Halq'eméylem language outside of learning situations such as classrooms. This does not seem to be related to any digital divide issues as participants confirmed that they and other community members use ICT on a regular basis for non-language related activities. There may be multiple reasons why participants do not use ICT with the Halq'eméylem language; however, Burton suggested that the primary reason that Halq'eméylem specific ICT is not used by community members is because people do not use the language:

> *"But the thing about the technology, the thing about everything, classes, education, language planning, everything that we try to do, it's like we're trying to support something that's not happening. So, if people were talking to their aunt and their grandmother or a couple that said we're going to make Halq'eméylem a part of our life and so on, then the technology and the classes could help them. But if all you have is the technology, then that's not going to solve the problem. The problem is a social problem, or a personal problem."* (personal communication, July 11, 2013)

Technology will be most successful where the language is already being used and, where the language is not being used, technology will not increase the usage as in the case of the Upriver Halq'eméylem language community. Technology to revitalize languages needs to be carefully planned with corresponding plans to introduce the technology and then post-implementation activities and oversight to ensure that the language continues to grow.

Bibliography

Adegbija, E. (2008). "Saving threatened languages in Africa: A case study of Oko." In Fishman, J.A. (ed.), *Can threatened languages be saved? Reversing language shift, revisited: A 21st century perspective*. Clevedon, UK: Multilingual Matters Ltd, pp. 284–308

Adley-Santa Maria, B. (1997). "White Mountain Apache language: Issues in language shift, textbook development, and native speaker-university collaboration." In Reyhner, J. (ed.), *Teaching indigenous languages*. Flagstaff, AZ: Northern Arizona University.

Battiste, M. (2008). "Research ethics for protecting indigenous knowledge and heritage." In Denzin, N.K., Lincoln, Y. S., and Smith, L. T. (eds.), *Handbook of critical and indigenous methodologies*. Thousand Oaks, CA: Sage Publications, pp. 497–509.

Berkes, F. (1993). "Traditional ecological knowledge in perspective." In Inglis, J. T. (ed.), *Traditional ecological knowledge: concepts and cases*. Ottawa, ON: IDRC/CRDI, pp. 1–9.

Cazden, C. B. (2003). "Sustaining Indigenous languages in cyberspace." In Reyhner, J., Trujillo, O., Carrasco, R. L., and Lockard, L. (eds.), *Nurturing native languages.* Flagstaff, AZ: Northern Arizona University, pp. 53–57.

Dauenhauer, N. M., and Dauenhauer, R. (1998). "Technical, emotional, and ideological issues in reversing language shift: Examples from Southeast Alaska." In Grenoble, L. A. and Whaley, L. J. (eds.), *Endangered languages: Language loss and community response*. Cambridge, MA: Cambridge University Press, pp. 57–98.

Eisenlohr, P. (2004). "Language revitalization and new technologies: Cultures of electronic mediation and the refiguring of communities." *Annual Review of Anthropology*, *33*:21–45.

First Peoples' Heritage, Language and Culture Council. (2010). *Report on the status of B.C. First Nations languages 2010*. Retrieved from http://www.fpcc.ca/files/PDF/2010-report-on-the-status-of-bc-first-nations-languages.pdf

Grenoble, L. A. (2011). "Language ecology and endangerment." In Austin, P. K. and Sallabank, J., *The Cambridge handbook of endangered languages*. Cambridge, MA: Cambridge University Press, pp. 27–44

Harrison, K. D. (2007). *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford: Oxford University Press.

Hinton, L. (2008a). "Language revitalization: An overview." In Hinton, L. and Hale, K. (eds.), *The green book of language revitalization in practice*. Bingley, UK: Emerald Group Publishing Limited, pp. 3–18.

Hinton, L. (2008b). "Learning and Teaching Endangered Indigenous Languages." In Van Deusen-Scholl, N. and Hornberger, N. H. (eds.), *Encyclopedia of Language and Education.* Springer Science+Business Media LLC, pp. 157–67.

Hornberger, N. H., and King, K. A. (2008). "Reversing Quechua language shift in South America." In Fishman, J.A. (ed.), *Can threatened languages be saved? Reversing language shift, revisited: A 21st century perspective.* Clevedon, UK: Multilingual Matters Ltd, pp. 166–194.

Kipp, D. (2009). "Encouragement, guidance and lessons learned: 21 years in the trenches of Indigenous language revitalization." In Reyhner, J. and Lockard, L. (eds.), *Indigenous language revitalization: encouragement, guidance and lessons learned*. Flagstaff, AZ: Northern Arizona University, pp. 1–9.

Kirkness, V. J. (1998). "The critical state of aboriginal languages in Canada." *Canadian Journal of Native Education*, *22*(1):93–107.

Kroskrity, P. V. (2009). "Language renewal as sites of language ideological struggle: the need for 'ideological clarification.'" In Reyhner, J. and L. Lockard (eds.), *Indigenous language revitalization: encouragement, guidance and lessons learned*. Flagstaff, AZ: Northern Arizona University, pp. 71–83.

Makagon, D., and Neumann, M. (2008). In *Recording Culture: Audio Documentary and the Ethnographic Experience*. Thousand Oaks, CA: SAGE Publications.

Norris, M. J. (1998). "Canada's aboriginal languages." *Canadian Social Trends*, *51*:8–16.

Tyler, T. R. (2002). "Is the Internet changing social life? It seems the more things change, the more they stay the same." *Journal of Social Issues*, *58*(1):195–205.

Warschauer, M. (2004). *Technology and social inclusion: Rethinking the digital divide.* Cambridge, MA: MIT Press.

Makers by Mail: Providing Makers Technologies to All

Christina Boyles-Petersen
christina-boyles@uiowa.edu
University of Iowa, United States of America

Lindsay Kistler Mattock
lindsay-mattock@uiowa.edu
University of Iowa, United States of America

Andrew Boyles-Peterson
andrew-petersen@uiowa.edu
University of Iowa, United States of America

This paper explores the development of Makers by Mail, a model for a mobile makerspace targeting graduate and undergraduate Digital Humanities scholars. While makerspaces (fablabs, hackerspaces, creative spaces) have steadily developed in libraries, university campuses, and community spaces over the past few years, these centers of innovation are often rooted in a particular place. Even mobile makerspaces, like the DHMakerBus, reach a limited audience by requiring users to gather in a particular place at a specific time. In response, we propose a new model for mobile makerspaces that emphasizes prototyping as pedagogy. This method involves shipping small, physical computing technologies and instructions to users using flat-rate priority mail boxes. Designed to support experiential learning - learning through making - the makerspace focuses on the computer processes that go into making rather than the products produced by them. As such, it emphasizes the process of making itself, whether successful, failed, or flawed.

Using an online checkout system, these kits are available for the cost of shipping (and the replacement cost of any lost or damaged parts). Training is provided through print materials as well as through the teaching commons on the Makers by Mail website. The mobility and affordability of these kits creates a new user experience for those involved by making digital tools accessible to individuals in traditionally unreached groups—rural and socioeconomically challenged communities. This model, therefore, gives a diverse group of users the opportunity to access advanced digital tools and, in the process, engages participants in critical making and maker culture, allowing them to gain confidence in coding, building, and digital literacy.

In *Debates in the Digital Humanities* Alexander Reid suggest that most graduate students have had little exposure to digital technology during their undergraduate education, “enter[ing] his or her graduate education as a novice in regards to the digital" (357). His assessment applies to our broader community of Eastern Iowa. Our initial findings suggest that many of the students approach technology with anxiety due to their lack of experience with computer programming. To address our users’ lack of technological expertise, we are scaffolding projects using a variety of technologies, including: littleBits, Raspberry Pi, and Arduino boards. These tools were selected for their mobility, accessibility, and affordability. Each of these factors allows for implementation of the makerspace in unconventional venues—secondary schools, adult education classes, and community events. Additionally, projects developed for each of these tools can be scaffolded, gradually increasing participants’ comfort and literacy with digital toolsand technologies.

This paper will report on the first phase of development of the Makers by Mail project. Over the first year of development, Makers by Mail kits were used in a variety of settings, from public makerspace events, to conference workshops, and in classrooms across the University of Iowa. In this second year, we are pursuing partnerships with community partners outside of Iowa City and beyond the borders of the state, including the University of Kansas, Grinnell

College, and the Williamsburg Public Library. For this paper, we will focus on the use of the kits as part of coursework for students in the University of Iowa School of Library and Information Science and humanities students enrolled in the Public Digital Humanities Certificate. These three case studies will illustrate the pedagogical potential of the kits and demonstrate three different modes of engagement with the technology.

Case 1: Teaching Digital Literacy with Raspberry Pi

All students enrolled in the Library and Information Science MA program must complete Computing Foundations in their first term of study. This course is designed as an introduction to many of the technologies that the students will encounter and engage with during their course of study at SLIS. The course provides a broad introduction to many technologies that Digital Humanities scholars utilize as well: XML, HTML, Python, and LAMP server-based platforms. For this audience we utilized a flipped classroom, using Raspberry Pi computer kits for a series of guided projects. By the end of the term, students had a working LAMP server with self-hosted websites and had engaged with encoding documents in XML, transformed XML into HTML with XLST, and interrogated those documents using Python scripts. This kit was designed not only to familiarize students with the technologies they are likely to encounter in practice, but to build key digital competencies in coding and working at the command line - skills necessary for engaging with Digital Humanities methodologies as well.

Case 2: Digital Humanities Project Development with Arduino

The second case reports on the use of Arduino kits for students enrolled in the university’s Introduction to Digital Humanities seminar. This course is designed to introduce students to digital humanities methodology and requires that they propose, research, develop, and launch a digital project. In this class, students use Arduino kits to create a data visualization of their research projects. Although initially skeptical, students demonstrated excitement once they had the opportunity to experiment and play with the technologies. By the end of the term, they developed creative and functional Arduino projects on topics including book censorship, digital art, and Walt Whitman. In doing so, they acquired both literacy with physical computing technologies and new understandings of data visualization.

Case 3: Designing Makers by Mail Kits

Our final case illustrates the philosophy of Makers by Mail and offers insight into the design and development of our kits. As part of the Digital Environments course (an elective course for DH Certificate students), students were asked to work with humanities faculty from the University of Iowa and other community partners to develop kits for targeted audiences. Over the first few weeks students were introduced to the concepts of critical digital pedagogy, design thinking, and values in design. They were then asked to apply theory to their practice by prototyping and developing kits and lesson plans to be implemented in specific classrooms and included on the Makers by Mail website. These kits included two Digital Shakespeare kits, a sound surveillance kit to pair with Hamlet, as well as a scene mapping kit using Arduino boards.

Methodology

In each of these cases, we assessed the increased digital literacy, competency, and confidence of the students through a series of surveys, journaling exercises, and participant observation. At the beginning and end of the term, each student completed a survey that asked students to rate their experience and comfort level with a variety of technologies (from computers, to cell phones, makers tech, and wearables), along with a series of open-ended questions gauging expectations and general feelings about technology. In Computing Foundations, students were also given a pre- and post-questionnaire designed to test and gauge the students' digital literacy.

Throughout the term, students were asked to keep a process log and reflective journal; this metacognitive strategy provided an opportunity for students to reflect on their thinking and learning process and produced multiple data points to track student progress throughout the semester. Finally, all students were required to submit one final reflection at the end of term, summarizing their experience and the major takeaways from the course.

Conclusions

Regardless of the student's indicated level of experience at the beginning of the term, all students reported an increase in their comfort and ability to engage with digital technology, from physical computing technologies like the Arduino to more traditional technologies like the Raspberry Pi computers. We argue that the self-contained nature of the kits and the pedagogy built into the technologies afforded students opportunities to play, tinker, and experiment, and that this trial-and-error experimentation and prototyping allowed students to critically engage with digital technologies in a way that the black-boxed, pre-packaged devices that students typically engage with do not. Further, the mobility of the kits offered opportunities for students to gain a sense of ownership of the technologies, transporting the kits from the classroom to their personal spaces and working without the constraints of a place-bound makerspace.

The use of the kits not only affords opportunities for makers to build confidence creating and experimenting with digital tools but also provides students with opportunities to critically engage with technology and build digital literacy skills that address other key aspects of DH education, including: working in interdisciplinary teams, applying digital practices, managing projects, and explaining technology (Rockwell and Sinclair 182-183). Ultimately, we argue that this model can be emulated in other educational settings as a new model for DH pedagogy that is more accessible and more collaborative than traditional makerspaces.

Bibliography

Association of College and Research Libraries (n.d.). "Information Literacy Competency Standards for Higher Education." http://www.ala.org/acrl/standards/informationliteracycompetency.

Bowler, L. (2014). "Creativity Through 'Maker' Experiences and Design Thinking in the Education of Librarians." *Knowledge Quest* 42 no. 5 (May/June 2014): 58-61.

DHMakerBus: http://dhmakerbus.com/.

Gierdowski, D., and Reis, D. (2015). "The MobileMaker: An Experiment with a Mobile Makerspace." *Library Hi Tech* 33, no. 4: 480-496.

McGrath, L.,and Guglielmo, L. (2015). "Communities of Practice and Makerspaces: DMAC's Influence on Technological Professional Development and Teaching Multimodal Composing." *Computers and Composition* 36: 44-53.

Moorefield-Lang, H.M. (2015)"When Makerspaces Go Mobile: Case Studies of Transportable Maker Locations." *Library Hi Tech* 33, no. 4: 462-471.

Rehbein, M.,and Fritz, C. (2012). "Hands-On Teaching Digital Humanities," in *Digital Humanities Pedagogy: Practices, Principles, and Politics*, 47-78, ed. Brett D. Hirsch. Cambridge, UK: Open Book Publishers.

Reid, A. (2012). "Graduate Education and the Ethics of the Digital Humanities," in *Debates in the Digital Humanities*, 350-367, ed. Matthew K. Gold. Minneapolis: University of Minnesota Press.

Rockwell, G., and Sinclair, S. (2012). "Acculturation and the Digital Humanities Community," in *Digital Humanities Pedagogy: Practices, Principles, and Politics*, 177-211, ed. Brett D. Hirsch. Cambridge, UK: Open Book Publishers.

Sayers, J., Elliott, D., Kraus, K.,et. al. (2016)"Between Bits and Atoms: Physical Computing and Desktop Fabrication in the Humanities," in *A New Companion to Digital Humanities*, 3-21, eds. Susan Schreibman, Ray Siemens, and John Unsworth. Malden, MA: Wiley Blackwell, 2016.

Unsworth, J. (2016) "4CAST '16 Keynote." Annual Campus Academic Strategies and Technology (4CAST) Conference, The University of Iowa, January 14, 2016.

UVic MakerLab: http://maker.uvic.ca/.

Scaffolded Hermeneutica for Literary Scholars with Novice Technical Skills

Jeremy Browne
jeremy_browne@byu.edu
Brigham Young University, United States of America

Hermeneutica

In *Hermeneutica*, Geoffrey Rockwell and Stéfan Sinclair (2016) argue for an approach to the digital humanities that deemphasizes the tool and positivist notions of proof. Their proposed approach, also called **Hermeneutica**, champions tool accessibility over tool sophistication. Similarly, scholarly play is legitimated as a useful step in developing research questions and as a means to reconsider established notions within literary disciplines. The aim of Hermeneutica as a methodology seems to be the generation of interesting humanistic questions as much as the resolution of open questions.

Rockwell and Sinclair demonstrate the difference between Hermeneutica and typical DH approaches by quoting from Gary Wong's 2009 blog post:

> [Typical DH] takes the worst part of the scientific papers (really really long sets of tabular data in the body of the text) and the worst part of papers from the humanities (really really complicated language where simple language would have done) and puts it in one. If this is what the cooperation of computational text analysis and traditional literary analysis yield, I am scared.

Because Hermeneutica attempts to join the **best parts** of these fields, it has the potential to turn DH into a discipline that is more useful for the vast majority of non-DH humanists. It could be the means of accelerating the mainstreaming of DH methods and bringing us to the eventual point where all humanities are digital—a destination Claire Clivaz described succinctly (DARIAH, 2016).

Voyant

One feature that distinguishes Hermeneutica from many other DH approaches is its companion set of tools meant to demonstrate its application. Voyant Tools, now referred to simply as Voyant, is a web-based, modular suite of tools meant to be "worth thinking **with**" (Rockwell and Sinclair, 2016: 10, original emphasis). The goal is to accommodate playful exploration of text and sharing of corpora across the web. It is not designed as an industrial-grade text analysis tool, but as a "toy" that allows scholars to uncover new questions and gain new appreciation of texts.

Current limitations of Hermeneutica

A fundamental component of Hermeneutica is that the scholar views text through the lens of Voyant (or other computational text analysis tools), and then synthesizes that experience with their prior knowledge of the text and its milieu. A problem that Voyant addresses, but does not solve, is that many scholars who know the most about specific texts lack the technological skills that would be considered pre-novice in DH circles. Voyant allows everyone with a text and a browser to explore word frequencies, collocations, etc., but it presupposes that the text is available and clean enough for use. In order for Hermeneutica to appeal to non-DH humanities scholars, these issues of text availability and the lack of user skill must first be addressed.

On the issue of text availability, it is not often that scholars wish to analyze text that is rare or missing. More often they are interested in text that is protected by various copyright laws, which prohibit posting the text to public websites such as Voyant. Thankfully, in the Unites States at least, Google Books' recent court victory (Stohr, 2016) now permits scholars to publish online the analysis results derived from copyrighted texts, so long as the original text is not recoverable by the user. To this end Rockwell and Sinclair developed Voyant 2's "non-consumptive" mode which restricts access to tools that allow full-text views.

While such developments represent Rockwell and Sinclair's amenability to meet the ever-evolving needs of Hermeneuticans, accommodating users' lack of technology skill is beyond the scope of their involvement. For example, it is not reasonable to expect the Voyant developers to be concerned over issues of text acquisition or text preparation. Rather, those concerns—while critical to expanding the pool of potential Hermeneuticans—are issues of local implementation. Similarly, it makes sense that Voyant would offer the ability to link to a corpus after uploading the text, but uploading the text and keeping track of various versions of corpora is beyond the scope of Voyant. A local practice of adding some structure around the Voyant suite ought to make Hermeneutica useful to a far greater audience than it is now.

Scaffolding

In the field of instructional design, such structure is called **scaffolding**. Specifically, scaffolding refers to the process of providing learners adequate introduction and examples before allowing them to attempt a task on their own (Bruner, 1978). For **scaffolded Hermeneutica**, DH-savvy professionals can work to acquire, clean, and upload text to Voyant (and other tools), and then provide public listings of the resulting corpora.

Examples of scaffolded Hermeneutica

We have implemented this scaffolded Hermeneutica approach in our Office of Digital Humanities beginning with the Cormac McCarthy Corpus Project (CMCP). The CMCP includes 13 Voyant corpora of McCarthy's 10 novels: one for the complete works, one for each novel, and two for novels (*The Orchard Keeper* and *The Road*) where the narration has been segregated from the dialogue. But the linchpin of scaffolded Hermeneutica is the CMCP's publicly-accessible website that organizes these Voyant corpora. The website is built on WordPress with the Pods content management plugin, and contains information on McCarthy's work, descriptions of Voyant (and other tools), and listings of links to the Voyant corpora. An essential feature of the website's structure is the ability to accommodate revisions to the current corpora as well as the addition of other tools in the future. Already, there is a non-Voyant sentence structure search tool attached as a beta-testing option.

A rough version of the Cormac McCarthy Corpus Project was presented at the 2015 conference of the Cormac McCarthy Society. The reaction to these tools being available for public use was strongly positive. One attendee referred to the website as "a game-changer."

The same scaffolded Hermeneutica is being implemented on two other projects: *Machado à longa distanciâ* and The Modernist Short Fiction Project. Preliminary demonstrations of the approach have yielded similar reactions to what we observed with the CMCP. Non-DH scholars become excited rather than anxious when the digital analysis tools are scaffolded to provide them ready access. In fact, these demonstrations turn into play sessions where non-DH scholars repeatedly request for certain words to be added to the frequency charts and other Voyant panels.

Conclusion

Hermeneutica and Voyant represent the greatest potential for growth in DH not because they are the most technologically or theoretically **advanced** developments, but because they are the most **accessible** to non-DH scholars. Still, they don't quite reach the ground level of technology skills possessed by most researchers in the humanities. The scaffolded Hermeneutica approach proposed in this paper seems to span that gap to make Hermeneutica more accessible.

Bibliography

Bruner, J. S. (1978). "The role of dialogue in language acquisition." In Sinclair, A., Jarvelle, R., J., and W. J.M. Levelt (eds), *The Child's Concept of Language.* New York: Springer-Verlag.

DARIAH (2016). My Digital Humanities – Part 1. YouTube. https://www.youtube.com/watch?v=I8aRtHW3b6g (accessed 1 November 2016).

Rockwell, G. and Sinclair, S. (2016). *Hermeneutica.* Cambridge: MIT Press.

Stohr, G. (2016). Google Book Project Can Proceed as Supreme Court Spurns Appeal. Bloomberg Politics. http://www.bloomberg.com/politics/articles/2016-04-18/google-book-project-can-proceed-as-top-u-s-court-spurns-appeal (accessed 1 November 2016).

Puns and Intertextuality: A Digital Approach to Greek Wordplay in Latin Literature

Evan Brubaker
ebrubak@tulane.edu
Tulane University, United States of America

Brandon LaFreniere
blaf@startmail.com
Tulane University, United States of America

Background

As one of the defining features of Latin literature, the influence exerted by the Greek linguistic and textual traditions has remained a key focus of the Classics. One of the manifestations of this intersection of cultures is in the topic of wordplay, or puns. Authors such as Vergil and Lucretius have been shown to utilize the metrical rules of pronunciation to embed Greek word forms in their poetry, and in the process, reference others works of literature and convey ideas beyond the superficial level of the text. While the influence of Greek writings upon Latin literature has received considerable scrutiny, the use of Greek puns has received only limited inquiry. Previous studies on puns, notably those of Snyder (1980) and Ahl (1985) are limited in scope and focus mainly on the relationship of Latin puns to Latin literature. Furthermore, studies which focus on Greek puns within Latin are limited primarily to writing about specific examples or a specific Latin author, such as O'Hara (1996) and his discussion of Greek word use in Vergil. Similarly, the recent (2013) compilation of studies edited by Kwapisz, Szymanski, and Petrain examining Greek and Latin wordplay does not look at the intersection of the two. Building upon previous strategies for data mining including the Tesserae Project, this paper seeks to offer a digital means to cross-reference Latin literature with Greek texts in order to find puns.

The program

Since the formation of our collaboration, we have produced software capable of detecting potential candidates for Greek puns. To achieve this end, we first set out to account for the rules of pronunciation within Latin poetry by referencing works such as Halporn et al. (1980), with issues such as elision and the lack of pronunciation of certain syllables addressed within the software. We then looked to set equivalence of pronunciation between Greek and Latin letters and groups of syllables, such as diphthongs. This was accomplished through examining transliterated Greek names and words which appear in Latin literature and using grammars including Smyth (1920). Additionally, certain Greek vowels and consonants which could correspond to multiple vowels, consonants, or syllables in Latin were also considered.

Accounting for these rules, we developed a processing system whereby Greek pronunciation equivalence is run through strands of Latin text. The software begins by reading one chosen Latin text in .txt format and then reading one or more Greek texts in .txt format, with all texts requiring UNICODE encoding for proper registration within the identification system. Once a Latin and Greek text(s) have been selected, the file paths are passed asynchronously into the model for processing. Both texts are then processed at the same time on different threads and the calling method waits for the completion of both methods.

The Greek text string is first processed by stripping off any accent or breathing mark and returning the base letter to lower case. While iterating through all characters in the text, when a space or new line character is reached, the previous word captured is added to a list. A word is not added to the list if it has previously been added to the list or the length of the word is less than a number for which the program allows specification. The result of this method is a list of those unique Greek words which appear in a chosen Greek text.

The Latin text string is first processed by converting all letters to lower case and taking account of the rules of elision and pronunciation. Similar to the Greek processing, the processed Latin text is converted into words. Each Latin word in the list is processed to determine if part of the word should be elided, based on the starting letter of the next word. The final step is to condense all the Latin words together into one text string without spaces, while keeping a list of the starting position of each word in the text string.

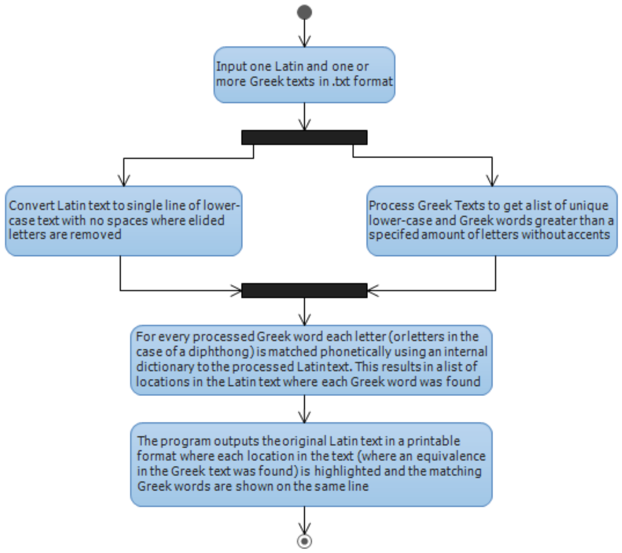
Once the two processing methods reach completion, the model has the list of unique Greek words and a string of Latin words, including the starting position in the original text of each word. The final processing method compares the Greek words to the Latin text string to find matches and returns a list of matches, which include the starting position of the match, ending position of the match, and the Greek word that matched. First, each letter of every Greek word is converted to a list of possible Latin characters or pairs of characters. A dictionary exists that for each base Greek lower case letter there corresponds a list of Latin letters or pairs of letters. Similarly, the existence of a diphthong is determined by referencing the values of a Greek-Latin diphthong dictionary. Both dictionaries were developed by examining those grammars and references noted earlier. By iterating though every Greek letter in a Greek word the possible Latin letters can be looked up in the dictionary, and, if the current letter and next letter in the Greek word are in the diphthong dictionary, this will override the regular dictionary lookup function. Once a list of every possible Latin correspondence for every letter in a Greek word exists, all permutations of letters are returned in a list of strings. This list is the list of every possible Latin series of characters that could match a Greek word. Using a regular expression every match in the Latin text string is found based on the previous list. These results are returned as a list of Greek words, including the starting and ending position in the original Latin text of the match.

The unique list of Latin-to-Greek results is returned to the view and is displayed in a grid, which can be sorted alphabetically, wherein the left column contains the uncompressed Latin character matches and the right column the Greek word equivalent in Greek characters. Also, a flow document is created which contains the original text with the matches highlighted in yellow and the matching Greek words to the right for each line. The Greek words in brackets are words that correspond to the same Latin text location, thus assisting the program user in contextualizing and applying significance to the results.

Results and applications

Through operation of the software, possible candidates for Greek puns within various Latin authors have been identified. By cross-referencing open-source texts of Vergil's *Aeneid* and the *Carmina* of Catullus with the Homeric epics and *Argonautica,* we have discovered hereto unnoticed Greek words, the meanings of which often correspond to the thematic concerns of the passage. Considering the limited extent of the search, both in duration and in those texts analyzed, these results seem to suggest that there are a considerable number of Greek puns remaining to be found within the corpus of Latin literature. While we are currently still in the process of refining the software and in consideration of more comprehensive ways to cross-reference Greek texts, we are still looking to expand the search both towards those texts already analyzed and to other Latin authors. With a view to the future, it is our hope that this software will eventually become available for public use and that it can lead to a more nuanced view of the role which Greek culture and language plays in Latin literature.

Program diagram


Input one Latin and one or more Greek texts in .txt format
Convert Latin text to single line of lower-case text with no spaces where elided letters are removed
Process Greek Texts to get a list of unique lower-case and Greek words greater than a specifed amount of letters without accents
For every processed Greek word each letter (or letters in the case of a diphthong) is matched phonetically using an internal dictionary to the processed Latin text. This results in a list of locations in the Latin text where each Greek word was found
The program outputs the original Latin text in a printable format where each location in the text (where an equivalence in the Greek text was found) is highlighted and the matching Greek words are shown on the same line

Bibliography

Halporn, J, et al. (1980). *The Meters of Greek and Latin Poetry*. Norman: University of Oklahoma Press.

Kwapisz, J, et al. (2013). *The Muse at Play: Riddles and Wordplay in Greek and Latin Poetry.* Berlin: De Gruyter.

O'Hara, J. (1996). *True Names: Vergil and the Alexandrian Tradition of Etymological Wordplay*. Ann Arbor: University of Michigan Press.

Smyth, H. (1920). *A Greek Grammar for Colleges.* New York: American Book Company.

Snyder, J. (1980). *Puns and Poetry in Lucretius' De Rerum Natura.* Amsterdam: Grul̀ner.

Layout analysis on newspaper archives

Vincent Buntinx
vincent.buntinx@epfl.ch
École Polytechnique Fédérale de Lausanne,
Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
École Polytechnique Fédérale de Lausanne,
Switzerland

Aris Xanthos
aris.xanthos@unil.ch
Université de Lausanne, Switzerland

The study of newspaper layout evolution through historical corpora has been addressed by diverse qualitative and quantitative methods in the past few years (Antonacopoulos et al, 2013; Gonzalez et al, 2001; Liu et al, 2001; Mitchell and Hong, 2004; Singh and Bhupendra, 2014). The recent availability of large corpora of newspapers is now making the quantitative analysis of layout evolution ever more popular. This research investigates a method for the automatic detection of layout evolution on scanned images with a factorial analysis approach. The notion of eigenpages is defined by analogy with eigenfaces used in face recognition processes. The corpus of scanned newspapers that was used contains 4 million press articles, covering about 200 years of archives. This method can automatically detect layout changes of a given newspaper over time, rebuilding a part of its past publishing strategy and retracing major changes in its history in terms of layout. Besides these advantages, it also makes it possible to compare several newspapers at the same time and therefore to compare the layout changes of multiple newspapers based only on scans of their issues.

Introduction to the Corpus

The corpus consists of digitized facsimiles of two Swiss newspapers, "Journal de Genève" (JDG) from years 1826 to 1997 and "Gazette de Lausanne" (GDL) from years 1804 to 1997. Scanned daily issues of each journal were transcribed using an optical character recognition (OCR) system (Rochat et al, 2016). The entire scanned data weighs more than 20TB, which makes most usual analysis techniques out of reach for regular desktop computers. This corpus has been the focus of several studies analyzing textual data (such as linguistic changes (Buntix et al, 2016) and named entity recognition (Ehrmann et al, 2016)). An example of different layouts of GDL's first page is given in Figure 1 which shows the evolution of various features, such as title size and position, fonts and number of columns.

Bitmap Factorial Analysis

In order to analyze layout evolution, we propose to build a static layout representation for every year in the corpus. Thus, when studying each newspaper's first page, we define the pixel t of the static representation $\overline{P_{y,m}}$ of month m of year y as

$$\overline{P^t_{y,m}} = \frac{1}{N_{ym}} \sum_{d=1}^{N_{ym}} P^t_{y,m,d}$$

Where N_{ym} is the number of issues in month m of year y and $P_{y,m,d}$ is the first page of day d of month m of year y. The pixel t of the static representation $\overline{\overline{P_y}}$ of year y is then defined as

$$\overline{\overline{P^t_y}} = \frac{1}{N_y} \sum_{m=1}^{N_y} \overline{P^t_{y,m}} = \frac{1}{N_y} \sum_{m=1}^{N_y} \frac{1}{N_{ym}} \sum_{d=1}^{N_{ym}} P^t_{y,m,d}$$

Were N_y is the number of month representations $\overline{P_{y,m}}$ in year y.

A diagram of the process is shown in figure 2.

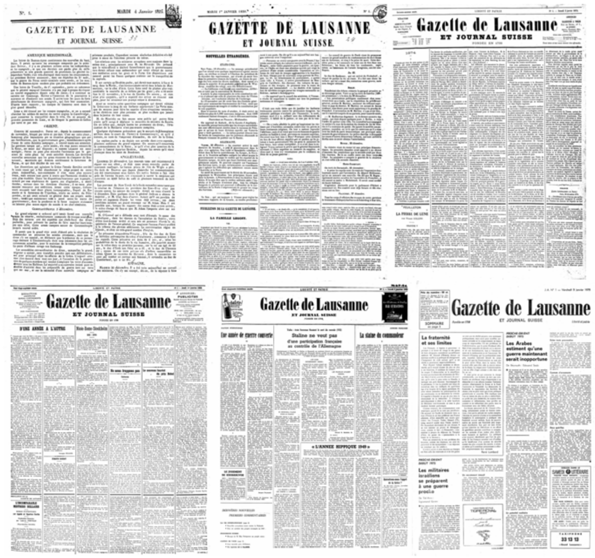


Figure 1. Different layouts of GDL in years 1825, 1850 and 1875 (top, left to right), 1925, 1950 and 1975 (bottom, left to right).

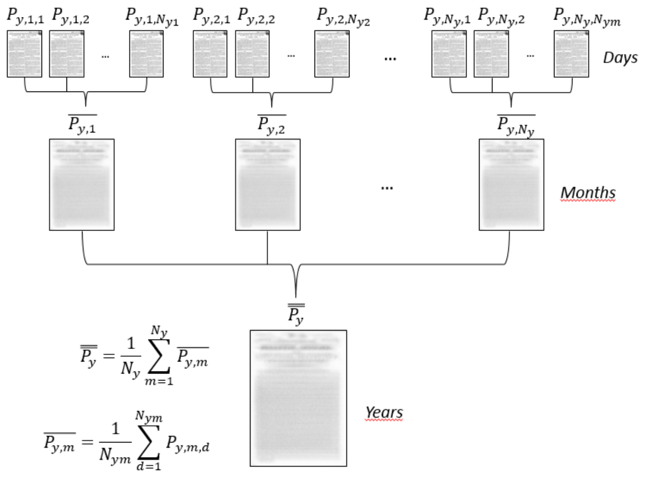

$P_{y,1,1}$ $P_{y,1,2}$ $P_{y,1,N_{y1}}$ $P_{y,2,1}$ $P_{y,2,2}$ $P_{y,2,N_{y2}}$ $P_{y,N_y,1}$ $P_{y,N_y,2}$ $P_{y,N_y,N_{ym}}$
Days
$\overline{P_{y,1}}$ $\overline{P_{y,2}}$ $\overline{P_{y,N_y}}$
Months
$\overline{\overline{P_y}}$
$\overline{\overline{P_y}} = \frac{1}{N_y}\sum_{m=1}^{N_y}\overline{P_{y,m}}$
$\overline{P_{y,m}} = \frac{1}{N_{ym}}\sum_{d=1}^{N_{ym}}P_{y,m,d}$
Years

Figure 2. Process diagram creating a yearly representation of first page layouts.

These representations give a vision of the mean layout over the course of a given year. Each yearly representation can be projected in a two-dimensional space by performing a principal component analysis (PCA) which maximizes the covariance on every pixel. This method is analogous to the eigenfaces method used for face recognition (Turk and Pentland, 1991a, 1991b) We compute the eigenvectors, that we named eigenpages, as well as the eigenvalues of the covariance matrix of the pixels. The yearly representations are then projected in the two-dimensional space of the two eigenvectors which have the highest eigenvalues. The resulting projections of yearly mean images of JDG and GDL from years 1900 to 1998 are portrayed in Figure 3. In these figures, each point is a yearly image and consecutive years are linked in order to highlight the change over time. The further apart the points are, the bigger the layout's changes occurring between two years. Visual inspection reveals several clusters of years with a similar layout. Furthermore, homogeneous sequences of years may be clustered automatically based on the (unprojected) distance between them (e.g. by computing the distance between year *y* and *y*+1 and "cutting" the sequence of years at positions where their distance exceeds an arbitrary threshold.

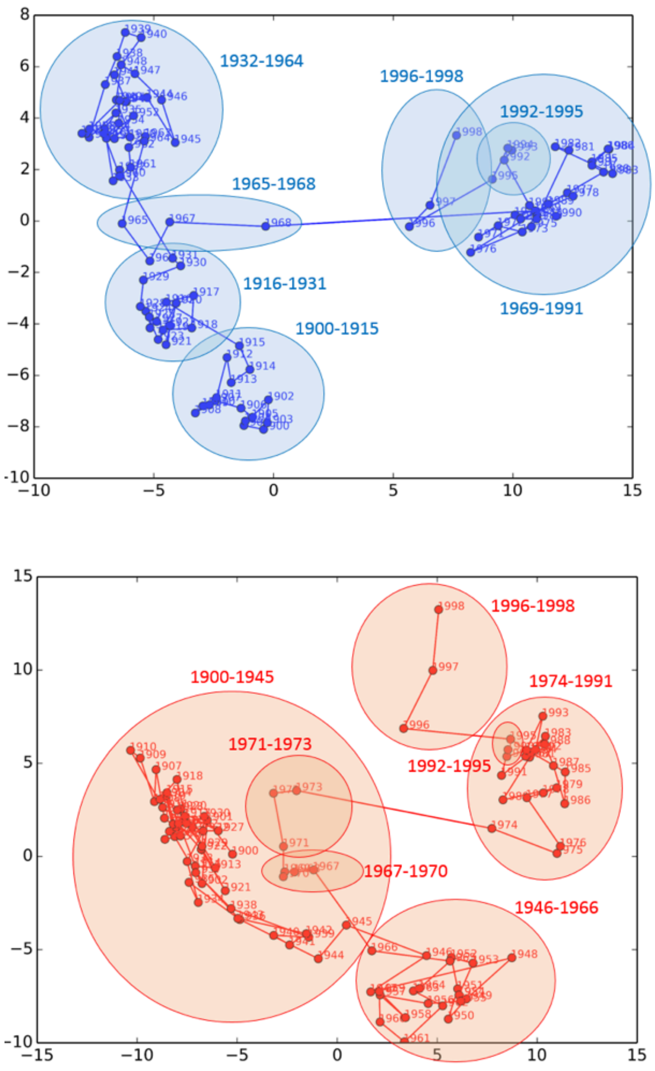


Figure 3: PCA projected results of the yearly representations of first pages of JDG (top, blue) and GDL (bottom, red) from years 1900 to 1998 with clusters obtained by visual inspection.

Discussion

The PCA technique allows us to quantify layout changes by covariance analysis of the pixels of yearly representations. The proportion of covariance information shown by the PCA is 73% for JDG and 76% for GDL. Visual interpretation reveals different chronological clusters which are displayed in Tables 1 and 2 along with their mean positions in the two-dimensional space of eigenpages as well as mean images representing these periods (computed in the same way as yearly images, cf. Figure 2). These mean images reveal the major layout transitions in each journal which may be summarized as follows:

Journal de Genève (JDG):

- 1900-1915: 6 columns, title above columns 2 to 5, little space between columns.
- 1916-1931: 4 columns, title above columns 1 to 4, more space between columns.
- 1932-1964: 4 columns, change of the layout around the title and the first title position.

- 1965-1968: 4 columns, change of the layout around the title, boxes with black borders begin to appear.
- 1969-1991: 4 columns, total change of the title, title above columns 2 to 4, logo appears, more space between columns and boxes, article titles are bigger.
- 1992-1995: 5 columns, fusion of JDG and GDL, big change of layout, boxes inside boxes begin to appear, more stable structure.
- 1996-1998: 6 columns, big change in title font, previous column layout replaced by a more classic one, article titles are placed at the top of the first page.

Gazette de Lausanne (GDL):

- 1900-1945: 6 columns, title above columns 2 to 5, little space between columns.
- 1946-1966: 7 columns, title above columns 2 to 6, more space between columns yielding particularly small column sizes.
- 1967-1970: 5 columns, title above columns 2 to 5, first column begins before the title which is on the right, advertisements placed below the page.
- 1971-1973: 6 columns, more classic layout with article titles at the top.
- 1974-1991: 4 columns, lots of space between columns and articles, bigger article titles.
- 1992-1995: 5 columns, fusion of JDG and GDL, big change of layout, boxes inside boxes begin to appear, more stable structure.
- 1996-1998: 6 columns, big change in title font, column layout replaced by a more classic one, the article titles are placed at the top of the first page.

The automatic clustering method described in previous chapter has been applied on unprojected distances and produce similar clustering results (depending on the threshold parameter). Qualitative analysis confirms that the resulting clusters are all separated by important layout transition phases.

Journal de Genève	1900 – 1915	1916 – 1931	1932 – 1964	1965 – 1968	1969 – 1991	1992 – 1995	1996 – 1998
Eigenpage 1	-1.5858	-4.5761	-6.3541	-4.0387	11.5680	9.6320	6.6127
Eigenpage 2	-6.9613	-3.4511	4.0575	-0.4756	0.9258	2.3969	1.2411

Table 1: Chronological clusters with their mean first page representations and their positions in the axes of PCA eigenpages (JDG). PCAPCgenpag(JDG)obtained by PCA for JDG.

Gazette de Lausanne	1900 – 1945	1946 – 1966	1967 – 1970	1971 – 1973	1974 – 1991	1992 – 1995	1996 – 1998
Eigenpage 1	-6.7016	4.5705	-2.1566	-2.6345	9.8505	9.5849	4.3915
Eigenpage 2	0.4252	-7.1833	-0.8661	2.4913	3.9961	6.3149	10.0302

Table 2: Chronological clusters with their mean first page representations and their positions in the axes of PCA eigenpages (GDL).

This analysis is also useful to compare several newspaper publishing strategies. We projected the two newspapers in the same two-dimensional space representation (presented in Figure 3) using the same method with yearly representations of both journals in order to compare their chronological trajectories. The covariance information shown by the PCA is 67%. Visual inspection reveals three main clusters for each journal. Each of these clusters turns out to correspond to groups of clusters that has been detected in the previous projections. We observe that the layout of both journals has evolved in a similar way but with different timescales. GDL is more dispersed than JDG and has explored different strategies during the period 1900-1966. However, GDL has adopted a style more similar to JDG style between 1967 and 1973 just before it entered a major layout transition in 1974 (5 years later than JDG).

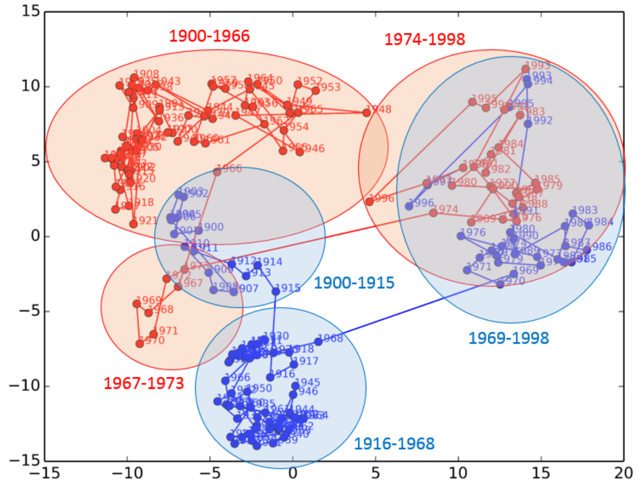


Figure 4. PCA projected results of the yearly representations of first pages of JDG (blue) and GDL (red) from years 1900 to 1998 in the same two-dimensional space representation with clusters obtained by visual inspection.

Conclusion

These first results demonstrate a promising method of detecting layout evolution automatically. The method is applicable to a large variety of longitudinal image corpora without any prerequisites, since it only requires images in bitmap format. It makes it possible to compare several corpora and determine periods of layout transitions in a common two-dimensional space for visual interpretation. In addition, unprojected distances can be used to determine layout changes in an entirely automatic fashion, by analyzing the representation space through clustering algorithms. Future work on this method should include the integration of an alignment method in the bitmap preprocessing step, because alignment errors may impact the pixel covariance analysis and eigenpages creation.

Bibliography

Antonacopoulos, A., Clausner, C., Papadopoulos, C., and Pletschacher, S. (2013) ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013. *12th International Conference on Document Analysis and Recognition*.

Buntinx, V., Bornet, C., and Kaplan, F. (2016) Studying Linguistic Changes on 200 Years of Newspapers. 2016. *DH2016*, Kraków, Poland, July 11-16.

Ehrmann, M., Colavizza, G., Rochat, Y., and Kaplan, F. (2016). Diachronic Evaluation of NER Systems on Old Newspapers. *13th Conference on Natural Language Processing (KONVENS 2016)*, Bochum, Germany, September 19-21.

González, J., Rojas, I., Pomares, H., Salmerón, M., Prieto, A., and Merelo, J.J. (2001) Optimization of web newspaper layout in real time. *Computer Networks*, Volume 36, Issues 2–3, July, Pages 311-321, ISSN 1389-1286, http://dx.doi.org/10.1016/S1389-1286(01)00158-X.

Liu, F., Luo, Y., Yoshikawaf, M., and Dongcheng, H. (2001). A New Component based Algorithm for Newspaper Layout Analysis. 2001. *6th International Conference on Document Analysis and Recognition*.

Mitchell, P. E., and Hong, Y. (2004) Newspaper layout analysis incorporating connected component separation. *Image and Vision Computing*, Volume 22, Issue 4, 1 April, Pages 307-317, ISSN 0262-8856, http://dx.doi.org/10.1016/j.imavis.2003.11.001.

Rochat, Y., Ehrmann, M., Buntinx, V., Bornet, C., and Kaplan, F. (2016). Navigating through 200 years of historical newspapers. 2016. *iPRES*, Bern, October 3-6.

Singh, V., and Bhupendra, K. (2014). Document layout analysis for Indian newspapers using contour based symbiotic approach. 2014. *International Conference on Computer Communication and Informatics (ICCCI-2014)*, Jan. 03 – 05, Coimbatore, INDIA

Turk. M., and Pentland, A. (1991a) Face recognition using eigenfaces. 1991. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. pp. 586–591.

Turk. M., and Pentland, A. (1991b) Eigenfaces for recognition. *Journal of Cognitive Neuroscience*. 3 (1): 71–86. doi:10.1162/jocn.1991.3.1.71. PMID 23964806.

La réception de l'édition critique numérique : accès multiples pour publics divers?

Joana Casenave
joana.casenave@umontreal.ca
Université de Montréal, Canada

Yves Marcoux
yves.marcoux@umontreal.ca
Université de Montréal, Canada

Introduction et cadre conceptuel

L'édition critique numérique, part active des Humanités numériques, bénéficie aujourd'hui d'un temps de recul suffisant pour dresser les premiers bilans des travaux édités sur le web. En effet, les premières éditions numériques ont aujourd'hui près de 20 ans : Peter Robinson a entrepris son projet d'édition des *Canterbury* Tales (*Les Contes de Canterbury*, au nombre de 24, ont été écrits par Geoffrey Chaucer au XIVe siècle. Le texte est écrit en moyen anglais et en vers) en 1993 pour en livrer une première version sur Cd-rom en 1996. De même, c'est en 1993 que Jerome McGann a lancé son projet *The Rossetti Archive*, avec l'objectif d'éditer l'ensemble de l'œuvre de l'écrivain anglais Dante Rossetti (1828-1882, peintre et écrivain britannique).

D'ailleurs, les éditeurs numériques eux-mêmes se plient régulièrement à l'exercice d'analyse de leur méthode (McGann, 2010 ; Robinson, 2004, 2010). Leurs réflexions sont nourries par leurs propres expériences et elles tirent également profit des études menées par les chercheurs spécialisés en Sciences de l'Information et en Humanités numériques qui s'appliquent, depuis quelques années, à théoriser l'édition critique numérique (Apollon *et al.,* 2014, Burdick *et al.*, 2012, Pierazzo, 2015, Sahle, 2013, Vanhoutte, 2010).

Au cours de cette présentation, nous allons précisément nous appuyer sur cette littérature analytique pour nous intéresser à la question de la réception et de la prise en compte de publics divers dans les éditions numériques.

Problématique

De fait, contrairement aux éditions critiques publiées sur papier, les éditions numériques semblent bénéficier d'un avantage certain : il s'agit de la possibilité offerte à l'éditeur de multiplier les voies d'accès au texte qu'il édite.

Dans une édition critique traditionnelle, publiée sur papier, l'éditeur propose une voie d'accès unique, qui prend forme dans son texte final établi, résultat de son interprétation critique des manuscrits témoins étudiés. Il adresse son édition à un public donné et travaille ensuite à rendre son étude conforme aux attentes de ces lecteurs préalablement identifiés. Pour qu'un même texte soit adressé à des publics divers, les éditeurs sont contraints de multiplier les éditions, ciblant un lectorat précis à travers chacune d'elles : les étudiants et leurs professeurs dans le cadre universitaire ; les chercheurs ; un public élargi, intéressé par l'édition savante des grandes œuvres.

Dans l'édition numérique, le format d'encodage XML/TEI permet de multiplier les niveaux d'information dans un même texte. L'éditeur peut ainsi coder des informations analytiques très diverses et choisir ensuite, de les présenter, ou non, aux lecteurs. Ce faisant, il peut décider de cibler plusieurs types de réception en adéquation avec les niveaux d'informations analytiques encodées. Il lui est alors possible de proposer les accès multiples et diversifiés qui répondent le mieux aux attentes des lecteurs ciblés.

Généralement, les projets d'éditions critiques numériques rassemblent, dès leur conception, des philologues pour établir le texte et des informaticiens pour s'occuper de la publication web et de l'élaboration de la plateforme de consultation. Ces équipes interdisciplinaires traitent à la fois les questions de philologie et les aspects liés à l'ergonomie et à l'« utilisabilité » de l'édition critique en cours de préparation. Désormais, dans le monde de l'édition numérique, le travail réalisé sur la réception de l'édition et l'accompagnement du lecteur est tout aussi important que la traditionnelle valorisation des sources philologiques.

La question qui nous occupe est donc celle de la prise en compte réelle de la diversité des publics dans l'édition numérique. Les outils techniques donnent – théoriquement du moins – la possibilité de multiplier les accès aux textes en fonction des lectorats. Qu'en est-il exactement ? Les éditeurs parviennent-ils réellement à multiplier les voies d'accès aux œuvres proposées ? Comment prennent-ils en compte, dans la préparation de la réception, la diversité des lecteurs et de leurs attentes ?

Méthodologie

L'édition numérique se situe pleinement dans le champ des Humanités numériques et, pour étudier cette question de la réception, nous allons adopter une méthodologie propre à ce champ disciplinaire, qui allie recherche et expérimentation.

Notre présentation se déroulera donc en deux temps. Nous allons tout d'abord mener une étude analytique sur un corpus d'éditions numériques sélectionnées à cet effet. L'objectif est de caractériser le positionnement des éditeurs numériques et d'observer les moyens qu'ils mettent en œuvre pour parvenir à cibler des publics différents sur une même plateforme de consultation des contenus, en proposant pour chacun d'eux des informations propres.

Ensuite, dans un deuxième temps nous présenterons les propositions concrètes que nous avons pu développer sur un petit corpus d'expérimentation. Au cours de la réflexion que nous avons menée sur une édition numérique d'un corpus de textes médiévaux, nous avons en effet cherché à multiplier les niveaux de réception et les modes de présentation des informations analytiques, en fonction de nos publics-cibles.

Résultats attendus

Partie I – analyses d'éditions numériques existantes

Pour étudier cette question, nous allons tout d'abord analyser un corpus d'éditions numériques sélectionnées par choix raisonné. Nous avons constitué un corpus de quatre éditions numériques qui se démarquent, dans le paysage éditorial contemporain, par l'attention portée, par les éditeurs, aux publics multiples. Les éditions critiques numériques retenues pour notre corpus d'observation sont diverses puisqu'elles portent sur des documents d'archives comme sur des œuvres littéraires.

Pour mener notre étude, nous avons élaboré une grille d'observation. Cette dernière est concentrée sur trois pôles : le texte édité, les outils de navigation et recherche, les outils de communication et de travail participatif proposés aux lecteurs. Les éditeurs y manifestent les particularités propres à chacun des publics cibles et y présentent des informations adaptées à chacun d'eux.

Les observations que nous ferons nous permettront alors de dégager des tendances et caractéristiques des éditions numériques afin de comprendre comment les éditeurs organisent, dans les plateformes de consultation de leurs éditions, les réceptions multiples.

Partie II – expérimentation sur un corpus littéraire médiéval

Dans un deuxième temps, nous allons confronter notre réflexion à notre propre travail d'expérimentation. La question des publics a été au centre de nos préoccupations lors de l'élaboration de notre édition numérique, et nous avons développé un schéma d'encodage permettant, précisément, de faire coexister des niveaux de réception multiples. Nous avons encodé et préparé la réception d'un ensemble divers d'informations analytiques : variantes philologiques, informations de critique littéraire, informations contextuelles. Ces informations sont alors présentées de manière coexistante mais distincte : certaines sont préparées en direction du public philologue spécialiste, d'autres en direction d'un public littéraire averti, d'autres en direction du grand public. Nous allons présenter notre édition actuellement en cours de réalisation, et expliciter la mise en place des niveaux de réception tels que nous les avons élaborés dans notre travail.

L'objectif de cette communication est ainsi de participer à une réflexion sur l'accès multiples aux textes ainsi que sur la prise en compte des niveaux de réception dans la préparation des éditions numériques.

Bibliographie

Apollon, D., Belisle, C., Régnier, P. (2014). *Digital Critical Editions*. University of Illinois Press.
Burdick, A. ***et al.*** *(*2012). *Digital humanities*. Cambridge: MIT Press.
Gold, M (dir.) (2012). *Debates in the Digital Humanities*. Minneapolis, University of Minnesota Press.
McGann, J. (2010). « Electronic Archives and Critical Editing », *Literature Compass*, vol. 7, nº 2, p. 37–42.
Pierazzo, E. (2015). *Digital Scholarly Editing*. Ashgate.
Robinson, P. (2010). « Editing Without Walls », *Literature Compass*, vol. 7, nº 2, p. 57–61.
Robinson, P. (2004). « ... but what kind of electronic editions should we be making? », University of Cologne.
Sahle, P. (2013). *Digitale Editionsformen : zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, Norderstedt, BoD, 3 vol.
Vanhoutte, E. (2010). Defining electronic editions: a historical and functional perspective. In *Text and Genre in Reconstruction. Effects of Digitalization on Ideas, Behaviours, Products and Institutions*. Cambridge: Open Book Publishers, pp. 119-144.
Warwick, C., Terras, M., Nyhan J. (2012). *Digital humanities in practice*. London: Facet Publishing in association with UCL Centre for Digital Humanities.

Digital musicology: through research and teaching

Tim Crawford
t.crawford@gold.ac.uk
Goldsmiths, University of London, United Kingdom

Kevin Page
kevin.page@oerc.ox.ac.uk
Oxford University, United Kingdom

David Lewis
david.lewis@oerc.ox.ac.uk
Oxford University, United Kingdom

David De Roure
david.deroure@oerc.ox.ac.uk
Oxford University, United Kingdom

Introduction

The Transforming Musicology project began in 2013, with funding from the UK's Arts and Humanities Research Council. It had three primary aims: to explore how technology could enhance, and perhaps 'transform', the practice and dissemination of conventional musicological research; to explore how large, new online sources of information, such as social media, might be exploited by digital musicology; and to look, in a modest way, at how digital approaches might be adopted in a sustainable way by the musicological community and beyond.

In this paper, we give an overview of the project and how it developed, discuss our strategy for approaching these goals, and reflect on how an inclusive strategy for both research and teaching can be effective, despite its organisational and financial costs.

The Transforming Musicology research project

Conventional musicology is represented in the project by two research strands, one drawing on music from the early modern period and with an emphasis on corpus-based research, and the other considering both the psychological perception and the historical reception of Wagner's use of leitmotifs, centred on his Ring cycle. The project also funded four mini-projects, selected from an open call, to extend the range of research that Transforming Musicology could cover. These, like our research on musicology of online communities, could then be treated as case studies for an analysis of technology needs and overlaps, points of connection, workflows and the areas where connective technologies such as Linked Data might be most effective.

The **early modern** strand of our project builds on two corpora: Early Music Online (EMO), scanned images and metadata for 300 sixteenth-century printed music books owned by the British Library; and the Electronic Corpus of Lute Music (ECOLM), a collection of full-text encodings and metadata for lute music. These contain closely interrelated music, and Optical Music Recognition tools were used to support the extraction of musical information from the facsimile images. Additional funding supported the creation of an open Linked Data union catalogue connecting the resources to one another and to relevant external resources, and the creation of a prototype application that supports both the linking process, and the publishing of web pages based on the now-linked resources. In collaboration with the BBC (the UK's national broadcaster) this resulted in demonstration pages for a BBC radio programme, the Early Music Show, allowing navigation of broadcast programme information, enriched with images and information from EMO, ECOLM, DBPedia and other sources of Linked Data (Weigl et al., 2016). Meanwhile, the musical content can be analysed to explore the creative process of arranging the vocal music of EMO for the lute, as seen in ECOLM (Lewis, Crawford & Müllensiefen, 2016)

The **Wagner** research strand has two elements. Firstly, we are exploring the early development of guidebooks and other media aimed at helping listeners recognise and follow the so-called leitmotifs in Wagner's four-opera cycle, Der Ring des Nibelungen. These diverge from the composer's own description of the use of motifs in his works, and tease out different aspects of the musical and dramatic themes of the works. In addition to the musicological question here, we also investigate whether a semantic web ontology can usefully represent the diversity of motifs, their different expressions and their relationships (Rindfleisch, 2016).

The second Wagner element is psychological, with biometric readings taken from audience members during performances of all four operas of the cycle. Along with memory tests, these provide us with a huge amount of data to interpret, and challenges for how to publish that data in ways that are useful to others. Further to this real-time physiological data, we also have expert annotations of a score, recording performance aspects that might be significant in the subject responses but not recorded in the notation, along with a complex set of information for aligning the time-based responses with each other and with a musical score (Baker & Müllensiefen, 2016).

A vast quantity of music of all kinds has become available on the web during recent decades and has engendered a correspondingly large amount of commentary, whether simply in the form of 'likes', or through intense online discussion on specialist websites to sophisticated scholarly articles. We carried out a pilot investigation of **user communities** on a lyric-annotation website, genius.com, formerly Rap Genius. With over 3,000,000 annotations, this can be regarded as a paradigmatic and valuable musicological resource which needs to be approached with the techniques of Big Data; our work has been centred on the overlaps between the networks of annotators and the songs and artists they annotate (Fields & Rhodes, 2016).

The four **mini-projects** have contributed: audio analysis of historical electronic music; ornamentation style in traditional Irish flute playing (Jančovič, Köküer, and Baptiste, 2015); a big data approach to finding the sources of the poetry used in medieval sacred songs; and an exploration of 18-20th-century London musical life based on digitised concert information from programmes and newspapers (Dix et al., 2014).

For these investigations, we have evaluated how **semantic web technologies** may offer solutions for bridging between disparate data and tool sets, and help document the research and its data, making it more reproducible as a result. (Nurmikko-Fuller and Page, 2016)

The Transforming Musicology website's Publications section offers a more complete list of research outputs than can be covered here.

We have embedded engagement with technology, musicology and other music-related communities in our work in general, but the most significant step for us in ensuring sustainability for digital musicology has been the creation of a week-long digital musicology workshop as part of the Digital Humanities at Oxford Summer School. The summer school is the largest of its kind in Europe, and is made up of a framework programme, consisting of a morning session each day, after which students attend the more specialist workshop they have selected.

By summer 2017, this will have run in three consecutive years and attracted a diverse set of over fifty students.

Teaching Digital Musicology at the Digital Humanities at Oxford Summer School

Although courses do exist on specialist areas within musicology (such as the Music Encoding Initiative summer school in Paderborn) this is the first dedicated to the whole discipline. In designing the curriculum, the Transforming Musicology team have reflected on how we can best make the methods that we have investigated over the course of the project – along with others that we know are used and work well – readily accessible to the wider musicology community.

Just as we have endeavoured to be inclusive in our approach to the musicological research that we have undertaken and supported within the project, we have also sought to create a space in which musicologists can be introduced to, and given the opportunity to experiment with, a wide variety of tools and approaches. It is important to us that the questions that musicologists investigate are not distorted when technology is brought to bear, and our teaching reflects this by balancing sessions about general-purpose tools with domain-specific use-case descriptions.

Students are eased into the week with an introduction giving an overview of the week ahead, but also some personal reflections about moving into a digital musicology approach from a more traditional background, and each day begins with presentations of case studies to motivate the techniques introduced during that day. Each day broadly considers a single topic, covering audio processing on both a small and large scale, music encoding, Optical Music Recognition and music processing. The final day presents more case studies and closes with a round table discussion which is intended to give the students the opportunity to reflect on the week and their own research practice and consider if and how the skills that they have learned can be useful to them.

The workshop is intended to stimulate ideas rather than to make programmers and audio engineers of our students, although we do still expect students to perform simple programming tasks and introduce them to key audio concepts and tools. Such lofty, long-term goals are difficult to evaluate. We take immediate feedback through a form on the last day, although the longer-term effect of attending can only be seen as the students develop their research in the coming years. Shorter-term impacts, including collaborations with tutors on papers and research proposals, and increasing contributions to conferences and workshops with an explicitly digital aspect can already be seen.

Musicology has always been interdisciplinary in nature, and has been transforming itself based on contributions from computing and web technology for over fifty years. We see our project as contributing to and supporting that transformation, both through our own research and through developing the skills of others.

Acknowledgements

Transforming Musicology is funded by the UK Arts & Humanities Research Council (ref: AH/L006820/1) and has benefitted from many contributors. The Transforming Musicology website lists the core researchers and investigators, the advisory panel, and mini-project

researchers and investigators. Similarly, there is too little space to acknowledge tutors, speakers and organisers for the summer school, but their contributions have informed the above discussion in many ways.

Bibliography

Weigl, D. M., Lewis, D., Crawford, T. and Knopke, I. (accepted) "On providing semantic alignment and unified access to music-library metadata." *International Journal of Digital Libraries*

Lewis, D., Crawford T. and Müllensiefen, D. (2016). "Instrumental Idiom in the 16th Century: Embellishment Patterns in Arrangements of Vocal Music." In *Proceedings of the International Society for Music Information Retrieval Conference*, New York, 524–530.

Rindfleisch, C. (2016). "'The eternal question to fate, surging up from the depth': Richard Wagner's Descriptions of his Leitmotives in Changing Contexts of Communication." *RMA Students' Conference* 2016, Bangor.

Baker, D. and Müllensiefen, D. (2016). "Hearing Wagner: Physiological Responses to Richard Wagner's Der Ring des Nibelungen." In *Proceedings of 14th International Conference for Music Cognition and Perception*, San Francisco, 207.

Fields, B. and Rhodes, C. (2016). "Listen To Me – Don't Listen To Me: What Communities of Critics Tell Us About Music." in *Proceedings of the International Society for Music Information Retrieval Conference*, New York, 199–205.

Jančovič, P., Köküer, M. and Baptiste, W. (2015). "Automatic transcription of ornamented Irish traditional flute music using hidden Markov models." In *Proceedings of the International Society for Music Information Retrieval Conference*, Malaga, 756–762.

Dix, A., Cowgill, R., Bashford, C., McVeigh, S. and Ridgewell, R. (2014). "Authority and Judgement in the Digital Archive" In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, ACM: New York, 1–8.

Nurmikko-Fuller, T. and Page, K. R. (2016). "A linked research network that is Transforming Musicology." In *Proceedings of the 1st Workshop on Humanities in the Semantic Web* co-located with 13th ESWC Conference, CEUR Vol. 1608, Aachen, 73–8.

The Extended Language of Religious Reform: Marking Up a Register for Early Modern Sermons

Thomas Winn Dabbs
tdabbs@cl.aoyama.ac.jp
Aoyama Gakuin University, Tokyo, Japan

This talk will report on the formation of an online open source register for early modern English sermons, a collaborative effort that includes scholars and technical experts from North America, the United Kingdom, and Japan. The need for a digital register of early modern sermons is abundantly clear to scholars who specialize in the history of the church and religious practice. Sermons were, for better or worse, central to efforts to broadcast and enforce disquieting religious reforms and to address the unsettling controversies in religious practice during the early modern period. It should be stressed, also, that a well-constructed and user-friendly platform of this nature would reach the many among the general public who are interested in the history of religion in general and in Christian church history in particular.

The term register is used here to denote a comprehensive and searchable list of English sermons, sermon events, and related sermon information from a variety of contemporaneous records and sources during the period 1500 to 1700, including sermons delivered to English audiences in Latin will also be catalogued. This project will not attempt to digitalize full texts of print or manuscript sermons.

Extant sermons were often printed long after they were delivered and many remain separately catalogued in manuscript. Non-extant sermons are known about from disparate sources: chronicles, diaries, private correspondence, and records of church and state. This lack of cohesion creates something of a vacuum amidst a plethora of sources that make reference to preaching in general and also to individual sermons. The goal of this project, therefore, is to bring to a list of sermons the cataloguing and search functionality (on a much smaller scale) that the British Library Short Title Catalogue (ESTC) online currently affords printed works.

There have been efforts recently to explore and document sermons during this period in digital formats but with differing objectives. These efforts include digital platforms in early development such as the announcement to 'Gateway to Early Modern Manuscript Sermons' (GEMMS), as well as digital projects such as The *St Paul's Cathedral Project* (formerly the Virtual Paul's Cross Project) and John Foxe's *The Acts and Monuments Online*. Standard resources for this project are, of course, the *ESTC Online*, *EEBO-TCP*, and for later reprints of early modern sermons Internet Archive has become increasingly useful. A host of fine scholars have been drawn to the study of early modern sermons, but this field has lacked an annotated and comprehensive register of sermons and sermon events.

The developers of this project will use a TEI-compliant XML structure. Much of the platform can be encoded with standardized markups for such elements as date, speaker, and location. However, this project is faced with the task of identifying and defining elements for a register in a space between digital bibliography and online textual archives or corpora.

Metadata will have to be extended to include an abundant number of tags for such areas as theological debate (e.g. purgatory, Eucharist) and church practice (e.g. baptism, burial). An exhaustive list of meta-concepts peculiar to the religious attitudes and movements of the period will have to be constructed. Among the many examples, such meta-tags as "anti-theatre," "Anabaptist," and "recantation," will have to be identified, agreed upon, and listed.

In sum, the greatest challenge for this project is in developing a comprehensive taxonomy for its metadata. In the view of the project team, the best way to approach this challenge is to start small and to learn by doing. Phase 1 of this project, therefore, is to digitalize and annotate Millar Maclure's register of the sermons of Paul's Cross along with the supplement to this edition.

Paul's Cross, the outdoor preaching pulpit located on the northeast side of the St Paul's precinct in early modern London, will be treated as ground zero for the cataloguing of early modern sermons. Located in the commercial and cultural center of London, Paul's Cross was the most influential public site for preaching to large, non-elite audiences in England during tumultuous periods of religious reform. This one locale had an immeasurable impact on Christian practice and public opinion from the Henrican period until the years leading up to the Civil War.

As the project expands toward its goal of being a comprehensive annotated register for all early modern sermons, the TEI XML encoding and extensions will allow scaling to a much larger platform. During Phase 1 the metadata arrangement for developers will be clarified. Should problems or omissions arise, which they invariably do, the fix can be made before the project expands. Finally, given that TEI has become the standard for what will become a partner project, this encoding will be highly adaptable should it eventually be contributed to a digital archive repository or OAIS.

Bibliography

ProQuest (2017). Early English Books Online (EEBO-TCP). Web (Subscription). eebo.chadwyck.com.

British Library (n.d). English Short Title Catalogue (ESTC). Web. estc.bl.uk.

GEMMS. (n.d.)'Gateway to Early Modern Manuscript Sermons' (GEMMS). Web. gemmsproject.blogspot.com.

Internet Archive (n.d.). Web. archive.org.

Foxe, J. (n.d.) John Foxe's The Acts and Monuments Online Web. johnfoxe.org.

Maclure, M. (1958). The Paul's Cross Sermons, 1534-1642 (Toronto: Toronto UP, 1958). Revised and augmented by Peter Pauls and Jackson Campbell Boswell. Centre for Reformation and Renaissance Studies (Ottawa: Dovehouse Editions, 1989).

NC State University (n.d.) The St Paul's Cathedral Project (formerly the Virtual Paul's Cross Project) Web. https://vpcp.chass.ncsu.edu.

A Stylometric Study of Nicholas of Montiéramey's Authorship in Bernard of Clairvaux's *Sermones de Diversis*

Jeroen De Gussem
jedgusse.degussem@ugent.be
Universiteit Gent, Belgium

Case Study

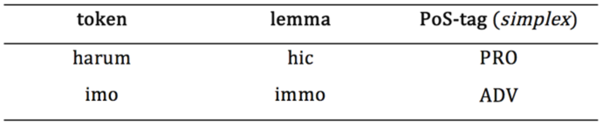
This short paper revisits the authorship of Bernard of Clairvaux's *Sermones de Diversis* (c. 1090 – 1153) through computational stylistics. Bernard's *De diversis* corpus comprises an assembly of unpolished and rudimentary sermons found in various, heterogeneous manuscripts. Bernard never disseminated the *De diversis* sermons himself, they have been first assembled, enumerated and published by Jean Mabillon in the 17th century (Callerot, 2006). Since Bernard of Clairvaux usually collaborated with secretaries, the obscure context of the corpus' composition and constitution has often made its sermons subject to some debate when it comes to Bernard's authorship. By 1145, the abbot's acclaim as the icon and figurehead of the Cistercian movement had brought along such a considerable administrative workload that the assistance of a group of secretaries was indispensable. These secretaries acted as Bernard's stand-ins, and spared him the time and effort it would cost of having to take up the quill himself at every single occasion. The *reportatio*, as it was called, entailed that the contents of Bernard's letters or sermons were engraved on wax tablets in a tachygraphic fashion. The cues, keywords and biblical references which Bernard had spoken aloud provided a framework that captured the gist of his diction. Afterwards, the scribe reconstructed what he had heard to a text on parchment which could pass for Bernard of Clairvaux's in its literary allure (Rassow, 1913; Leclercq, (2)1962; Constable, 1972). Amongst these amanuenses, Nicholas of Montiéramey († 1176 / 78) was as a focal figure, and a highly skilled imitator of his master's writing style. The influence of Nicholas' mediation on several particular text instances within Bernard's *De diversis*, and more generally on his entire oeuvre, has fallen subject to much debate.

Nicholas began serving Bernard as an emissary around 1138-41, carrying letters concerning Abelard's heresy to Rome (Turcan-Verkerk 2015). His literary qualities, likely to have been acquired through his education in the Benedictine abbey of Montiéramey, enabled him to immediately enter the scriptorium and officially become Bernard's closest secretary. Their friendship, however, knew an abrupt and painful ending in the final years of Bernard's life, around 1151-2, when Nicholas must have severely breached his master's trust by sending out letters without his permission. The scandal has for a long time upheld Nicholas' portrayal as a disreputable Judas by Bernard's side, a condemnation which has shimmered through on a scholarly level as well, and has resulted in highly subjective and speculative attributions. For instance, Nicholas has been found sending out Bernard's *De diversis* 6, 7, 21, 62, 83, 100 and 104 in a letter to count Henry the Liberal, claiming that they were "of [his] invention, of [his] style, aside from what was taken from others in a few places" (Leclercq, (1)1962). Also *De diversis* 40, 41 and 42 have been found within Nicholas' oeuvre (Rochais, 1962). Nevertheless, Nicholas' reputation as a fraud and a plagiarist has withheld 20th-century scholars such as Leclercq and Rochais to believe that his claim to authorship is any sense warranted, and has maintained the sermons' authenticity as uncontested, even despite the fact that later scholars have seriously doubted their views on Nicholas of Montiéramey's alleged deceitfulness and falsification (Jaeger, 1980; Constable, 1996). The temptation for historians to draw lines in between imitation and plagiarism in order to categorize writings and collate them in attributed editions, valuable as it is, can also be rather anachronistic or even unbefitting in a medieval context (Nichols, 1990; Cerquiglini, 1999). Perhaps Nicholas felt himself to be a rightful partaker in the composition of these works, a participation which might disclose itself stylistically.

Stylometry

The texts of Bernard of Clairvaux are edited in the *Sancti Bernardi Opera* (SBO), Jean Leclercq's edition published in 8 volumes. Nicholas' letters and sermons, on the other hand, still lack a modern edition, and can only be found in Migne's Patrologia Latina (see table 4). Although experiments and debates as to which textual features best capture stylistic difference are still ongoing, many state-of-the-art studies employ function words, which still prove to be the most robust discriminators for writing styles (Burrows, 2002). Function words are usually short and insignificant words that pass unnoticed, such as pronouns, auxiliary verbs, articles, conjunctions and particles, whose main advantage is their frequent occurrence, less conscious use by authors and content- or genre-independent character. Their benefit and success for stylometry in Latin prose have been convincingly demonstrated before, although the methodology still raises acute questions which keep stylometrists on the lookout for alternatives.

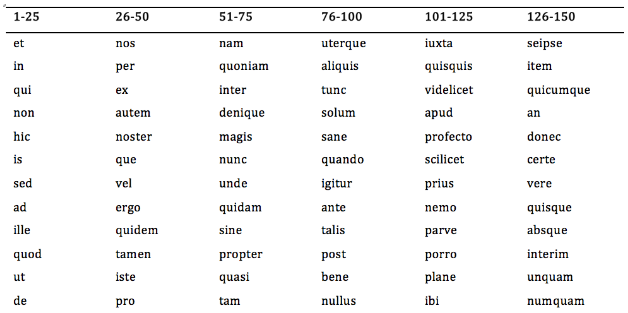
Because medieval Latin is a synthetic language with a high degree of inflection, the texts required some preprocessing (Mantello and Rigg, 1996). For instance, enclitica such as *-que* and *-ve* had to be separated from the token in order to be recognized as a feature. Secondly, texts are more easily mined for information when the lexemes are lemmatized (which means that the instance of the word is referred to its headword) and when its words (tokens) are classified according to grammatical categories (parts of speech). For this purpose we applied the Pandora lemmatizer-tagger on the texts, a piece of software developed by Kestemont and De Gussem that is equipped to achieve specifically this (Kestemont and De Gussem, forthcoming).



token	lemma	PoS-tag (*simplex*)
harum	hic	PRO
imo	immo	ADV

Figure 1. Excerpt from table showing tags applied to the texts

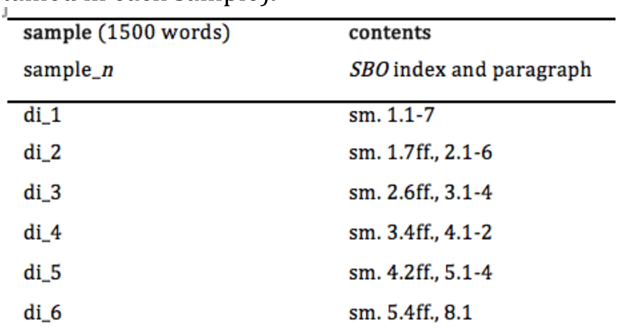
The third column, the part-of-speech-tag (PoS), allowed to immediately restrict the culling of most frequent words to those word categories that make up the collection of function words: conjunctions (CON), prepositions (AP), pronouns (PRO) and adverbs (ADV). This likewise filtered out some noise caused by ambiguities or homonyms like *secundum*. Once procedures of this sort were carried out in full, we arrived at a list of the 150 most frequent function words (MFFW) of the corpus (Figure 2)



1-25	26-50	51-75	76-100	101-125	126-150
et	nos	nam	uterque	iuxta	seipse
in	per	quoniam	aliquis	quisquis	item
qui	ex	inter	tunc	videlicet	quicumque
non	autem	denique	solum	apud	an
hic	noster	magis	sane	profecto	donec
is	que	nunc	quando	scilicet	certe
sed	vel	unde	igitur	prius	vere
ad	ergo	quidam	ante	nemo	quisque
ille	quidem	sine	talis	parve	absque
quod	tamen	propter	post	porro	interim
ut	iste	quasi	bene	plane	unquam
de	pro	tam	nullus	ibi	numquam

Figure 2. Excerpt from contents of a table showing most frequently occurring function words.

Each of the corpora was segmented into samples. This yields the advantage of "effectively [assessing] the internal stylistic coherence of works," (Eder et al., 2016) which answers directly to the primary goal of the present study. The sermons were segmented into 1500 word-samples (Figures 3-4 present aexcerpts from tables describing the texts contained in each sample).



sample (1500 words) sample_*n*	contents *SBO* index and paragraph
di_1	sm. 1.1-7
di_2	sm. 1.7ff., 2.1-6
di_3	sm. 2.6ff., 3.1-4
di_4	sm. 3.4ff., 4.1-2
di_5	sm. 4.2ff., 5.1-4
di_6	sm. 5.4ff., 8.1

Figure 3. Excerpt from a table describing the sample contents (1500 words) for Bernard's Sermones de Diversis as shown in figures 5-7.

sample (1500 words) sample_*n*	contents *PL* (vol: col.)
ep_1	ep. 1 (196: 1593a-1594b)
	ep. 2 (196: 1594b-1596a)
	ep. 3 (196: 1596b-1597b)
ep_2	ep. 4 (196: 1597b-1598c)
	ep. 5 (196: 1598d-1600a)
	ep. 6 (196: 1600b-1601b)

Figure 4. Excerpt from a table describing the sample contents (1500 words) for Nicholas' sermons and letters as shown in figures 5-7.

It should be noted that 1500 word-samples run the risk of increased imprecision, a consideration which should nuance any interpretation of the results (Kestemont et al., 2013). Once the corpus was divided, each of the text samples was vectorized to document vectors. The raw counts were TF-IDF-normalized (*Term frequency inverse document frequency*), a procedure which divides the function word frequencies by the amount of text samples that respective function word appears in (Manning, 2008; Kestemont et al., 2016). As a consequence, less common function words received a higher weight which prevents them from sinking away (and losing statistical significance) in between very common function words. Once the data was preprocessed and regulated, two statistical techniques were applied to visualize its dynamics.

The first is a *k* Nearest Neighbors network in GEPHI (hereafter abbreviated to *k*-NN) (Jockers, 2013; Eder, 2015; Jacomy et al., 2014), the second is principal components analysis (hereafter PCA) (Binongo et al., 1999). Their respective results will prove to be similar in a general sense, yet crucially different in the details. We argue that such an additional statistical validation provides for a more accurate, nuanced interpretation and a better intuition of the data. In the first visualization, the k-NN networks, we first calculated the 5 closest text samples to each text sample by applying k-NN on the frequency vectors. Accordingly for each text the 5 most similar or closest texts were calculated, weighted in rank of smallest pairwise distance (Minkowski metric, a Eucledian metric) and consequently mapped in space through force-directed graph drawing (algorithm Force Atlas 2). The weights were directly calculated from the distances. The intuition is then that the distances should be normalized to a (1,0) range (as a higher distance responds to a lower weight). Secondly, PCA is a technique that allows to reduce a multivariate or multidimensional dataset of many features, such as our function word frequencies, to merely 2 or 3 principal components which disregard inconsequential information or noise in the dataset and reveal its important dynamics. The assumption is that the main principal components, our axes in the plot, point in the direction of the most significant change in our data, so that clustering and outliers become clearly visible. Each word in our feature vector is assigned a weighting or loading, which reflects whether or not a word correlates highly with a PC and therefore gains importance as a discriminator in writing style. In a plot, the loadings or function words which overlap with the clustered texts of a particular author are the preferred function words of that author (see Figure 7 under "Results").

Results

Figure 5 (*k*-NN) and Figure 6 (PCA) feature the results of matching up Bernard's *Sermones de Diversis* benchmarked against the latter's *Sermones Super Cantica Canticorum* (his literary masterpiece) and the sermons and letters of his secretary Nicholas of Montiéramey.

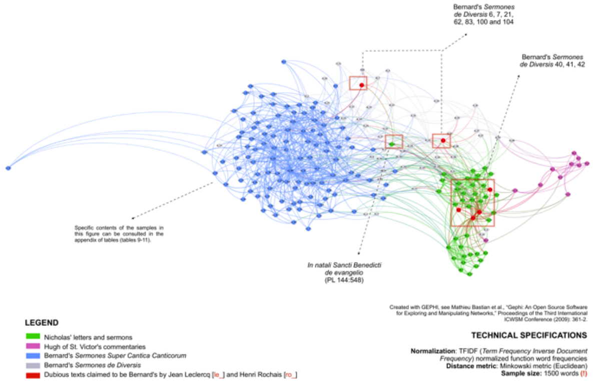


Figure 5: k-NN Networks

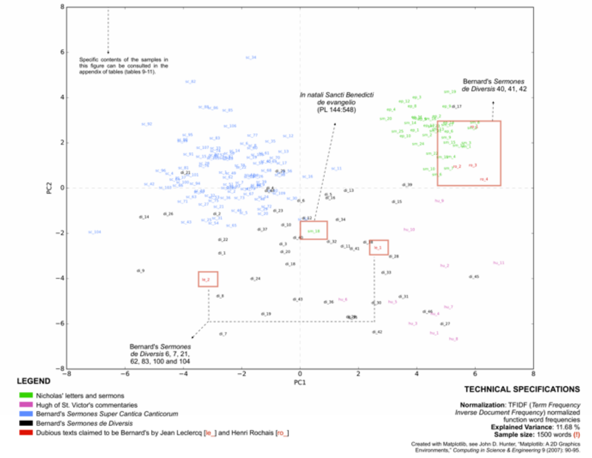
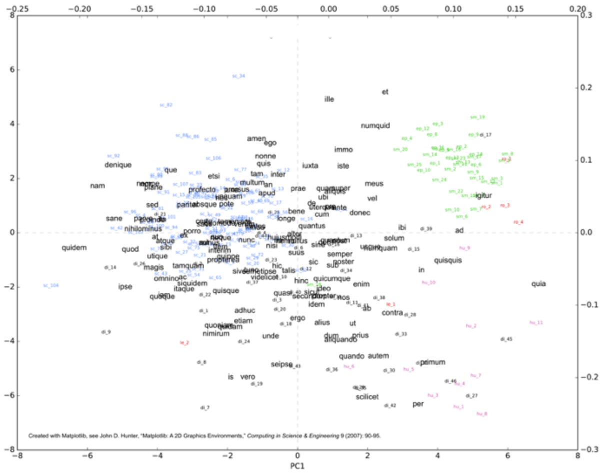


Figure 6. Principal Components Analysis (PCA)



Created with Matplotlib, see John D. Hunter, "Matplotlib: A 2D Graphics Environments," *Computing in Science & Engineering* 9 (2007): 90-95.

Figure 7. PCA Loadings

Firstly, when examining the visualizations, it is striking how – indeed – the diversity of Bernard's *De diversis* is captured. Especially PCA demonstrates the discernible stylistic incoherence as the samples burst open all over the plot (especially along the vertical axis of the second principal component), at times suggesting the interference of other writers than Nicholas or Bernard in their composition. Other samples gravitate in between Nicholas and Bernard, and in some cases Nicholas' influence on the style is undeniable. *De diversis* 6, 7, 21, 62, 83, 100 and 104, which Nicholas included in the letter to count Henry the Liberal (they are split up in two red samples labeled with le_ of Leclercq), do not betray an obvious affinity to Nicholas' style (although le_1 is not far off). Neither are they unambiguously Bernard's. Both samples diverge strongly and seem too hybrid in nature to be restrained. The case rather ostensifies how difficult it is to defend concepts such as "single authorship" or "text theft" in a medieval context: the le_ samples are clearly not of a "singular" style (nor of Nicholas's style, nor of Bernard's), but defy classification. In fact, if we compare both *k*-NN and PCA, Nicholas' influence in sample le_1 seems considerably larger than Bernard's. It has by now become an untenable simplification to argue that Nicholas has stolen these sermons, especially if we review the results of the second case, that of *De diversis* 40, 41 and 42 (four red samples labeled with ro_ of Rochais): although the sermons emanate from bernardian thought, PCA and *k*-NN unambiguously cluster all three sermons together with those written by Nicholas, not Bernard.

Bibliography

Binongo, J.N.G., and Smith, M.W.A. (1999). The Application of Principal Components Analysis to Stylometry, Literary and Linguistic Computing 14(4): 446-66.

Burrows, J.F. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing 17(3): 267-87.

Callerot, F. (2006). Introduction. In Leclercq, J., Rochais, H.M. and Talbot, C.H. (eds.), Bernard of Clairvaux, Sermons Divers (3 vols). Paris: Sources Chrétiennes, pp. 1:21-60

Cerquiglini, B. (1999). Wing, B. (transl.) In Praise of the Variant: A Critical History of Philology. Baltimore: John Hopkins University Press.

Constable, G. (1996). Forgery and Plagiarism in the Middle Ages. In Culture and Spirituality in Medieval Europe. Aldershot: Variorium, pp. 1-41.

Constable, G.. (1967). Nicholas of Montiéramey and Peter the Venerable. In The Letters of Peter the Venerable (2 vols.). London: Harvard University Press, pp. 2: 316-330.

Constable, G.. (1994). The Language of Preaching in the Twelfth Century, Viator 25: 131-52.

Eder, M. (2015). Visualization in Stylometry: Cluster Analysis Using Networks, Digital Scholarship in the Humanities: 1-15.

Eder M., Rybicki J. and Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis, The R Journal 16(1): 1-15.

Jacomy, M. et al. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, PLoS ONE 9(6): e98679. doi:10.1371/journal.pone.0098679 (accessed 6 April 2017).

Jaeger, S.C. (1980). Prologue to the Historia Calamitatum and the "Authenticity Question", Euphorion 74: 1-15.

Jockers, M.L. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.

Kestemont, M., Moens, S. and Deploige, J. (2013). Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux, Digital Scholarship in the Humanities 30(1): 199-224.

Kestemont, M. and De Gussem, J. (forthcoming). Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. Journal of Data Mining and Digital Humanities.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016). Authenticating the writings of Julius Caesar, Expert Systems with Applications 63: 86-96.

Leclercq, J. (1962a). Les collections de sermons de Nicolas de Clairvaux. In Recueil d'études sur saint Bernard et ses écrits (4 vols.). Rome: Edizioni di storia e letterature, pp. 1:47-82.

Leclercq, J. (1962b). Saint Bernard et ses secrétaires. In Recueil d'études sur saint Bernard et ses écrits (4 vols.). Rome: Edizioni di storia e letterature, pp. 1: 3-25.

Leclercq, J. (1969). Notes sur la tradition des épitres de S. Bernard Recueil d'études sur saint Bernard et ses écrits (4 vols.). Rome: Edizioni di storia e letterature, 3:307-22.

Mantello, F. A. C. and Rigg, A. G. (eds.) (1996). Medieval Latin: An Introduction and Bibliographical Guide. Washington (D.C.): Catholic University of America Press.

Nichols, S.G. (1990). Introduction: Philology in a Manuscript Culture, Speculum 65(1): 1-10.

Rassow, P. (1913). Die Kanzlei St. Bernhards von Clairvaux. Studien und Mitteilungen zur Geschichte des Benediktiner-Ordens 34. London: FB&c Ltd.

Rochais, H.M. (1962). Saint Bernard est-il l'auteur des sermons 40, 41 et 42 «De diversis»? Revue Bénédictine 72(3-4): 324-345.

Turcan-Verkerk, A.-M. (2015). L'introduction de l'ars dictaminis en France: Nicholas de Montiéramey, un professionel du dictamen entre 1140 et 1158. In Turcan-Verkerk A.-M. and Grévin B. (eds.) Le dictamen dans tous ses états: perspectives de recherche sur la théorie et la pratique de l'ars dictaminis (xie-xve siècles). Turnhout: Brepols, pp. 63-98.

Towards a Linked Data Access to Folktales classified by Thompson's Motifs and Aarne-Thompson-Uther's Types

Thierry Declerk
declerck@dfki.de
German Research Center for Artificial Intelligence Germany

Antónia Koštová
akostova@coli.uni-saarland.de
Saarland University, Germany

Lisa Schäfer
lisas@coli.uni-saarland.de
Saarland University, Germany

Introduction

We present in this paper work consisting in porting to an integrated ontology two central resources for the classification of folktales: The "Motif-index of folk-literature" (Thompson, 1977) and the "Types of International Folktales" (Uther, 2004). The first resource, often called Thompson Motif Index (abbreviated as TMI) is available as an online resource. The second resource is a classification system for folktale types, which was published by Hans-Jörg Uther (2004), extending former work by Antti Aarne (1961) and Stith Thompson (1977). In the following we are using the acronym ATU for referring to this classification system. Recently, a large amount of the ATU data has been made available online, offering also annotation facilities for tales in multilingual versions.

Our work consisted in extracting from those resources, which are stored in different formats, classification relevant information and re-organizing them in two interrelated ontologies, using for this the W3C standards OWL (which stands for "Web Ontology Language" , RDF(s)– see the W3C recommendations for more details– and RDF. The aim is to make those classification resources machine readable, interoperable and to support by this formal representation of the metadata access to folktales annotated by those classification systems in the context of the Linked Open Data framework.

Thompson Motif Index

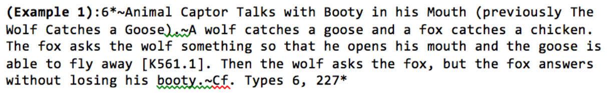
A folktale motif can be defined as a "repeated story element, e.g., a character, an object, an action, or an event that can be found in several stories". In TMI motifs are organized in a tree structure, providing for a parent-child relation between the listed elements. One motif entry consists of a motif-id and a motif name. Optionally, a **motif description** and **references** are provided. Table 1 provides for an example of few motifs illustrating the tree structure and hierarchy of TMI.

Motif-id	Motif name
A	Mythological motifs
A1	Identity of creator
A1.1	Sun-god as creator
A1.2	Grandfather as creator
A1.3	Stone-woman as creator
A1.4	Brahma as creator
A2	Multiple creators

Table 1: A few motifs from Motif-index of folk-literature and their hierarchical organisation

Aarne-Thompson-Uther Folktales Types (ATU)

A folktale type can be described as a main story line that can be found in several cultures. The parts of this story line can refer to specific story elements also known as motifs. A folktale type is therefore a bigger unit than a motif. As can be seen in example 1, an entry in the ATU system consists of a type id ("6*"), a title ("Animal Captor Talks with Booty in his Mouth (previously The Wolf Catches a Goose).") and a text summarizing the typical "storyline" of this type of folktale. Within or at the end of this "script", links to corresponding Thompson Motif-Indices can be provided ("[K561.1]"). Finally (and optionally), similar or related types can be indicated.

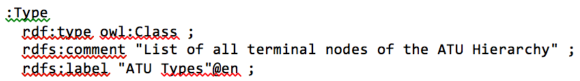


```
(Example 1):6*~Animal Captor Talks with Booty in his Mouth (previously The
Wolf Catches a Goose).~A wolf catches a goose and a fox catches a chicken.
The fox asks the wolf something so that he opens his mouth and the goose is
able to fly away [K561.1]. Then the wolf asks the fox, but the fox answers
without losing his booty.~Cf. Types 6, 227*
```

Generation of the Ontology

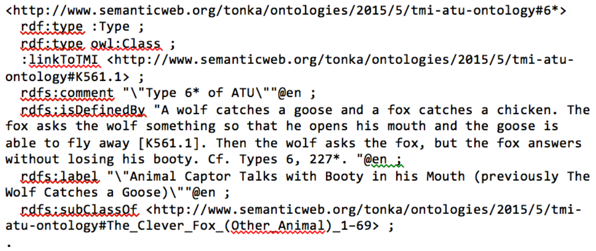
The OWL and RDF(s) representation for the ontology was generated semi-automatically from the html code of both TMI and ATU, responding to few design decisions we had to take. For TMI we went for a double representation: the hierarchy structure of the IDs is represented as an OWL subclass hierarchy, but all terminal nodes (leaves of the tree) are represented as both an instance of a class we call "Motif" and as an instance of the pre-terminal node in the taxonomy. This reflects our intuition that what Thompson called a motif is in most of the cases the content of the terminal nodes of the classification system, while the non-terminal nodes are more to be considered as abstraction helping in the taxonomic structures.

We compared the automatically created ATU part of the ontology to the printed version of Hans-Jörg Uther's "The Types of International Folktales". Using the ontology editor "Protégé", we manually added missing subclasses and individuals, rearranged generated classes and corrected errors such as typos in the electronic version of ATU or splitting errors because of inconsistent punctuation. By this step we obtained 2802 ATU classes organized in seven main subclasses, which have also subclasses, in accordance with the hierarchical structure of types proposed in (Uther, 2004). Below we display examples of the encoding of ATU data in our ontology. We first show a main class (we use the Turtle syntax for serializing the RDF code) of our ATU model ":Type":



```
:Type
  rdf:type owl:Class ;
  rdfs:comment "List of all terminal nodes of the ATU Hierarchy" ;
  rdfs:label "ATU Types"@en ;
```

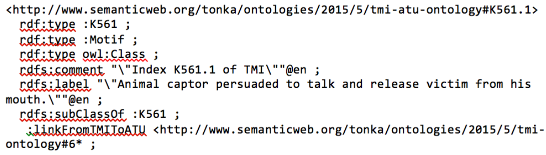
A subclass of "Type", for example the type "6*", has the following syntax:



```
<http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-atu-ontology#6*>
  rdf:type :Type ;
  rdf:type owl:Class ;
  :linkToTMI <http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-atu-
ontology#K561.1> ;
  rdfs:comment "\"Type 6* of ATU\""@en ;
  rdfs:isDefinedBy "A wolf catches a goose and a fox catches a chicken. The
fox asks the wolf something so that he opens his mouth and the goose is
able to fly away [K561.1]. Then the wolf asks the fox, but the fox answers
without losing his booty. Cf. Types 6, 227*. "@en ;
  rdfs:label "\"Animal Captor Talks with Booty in his Mouth (previously The
Wolf Catches a Goose)\""@en ;
  rdfs:subClassOf <http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-
atu-ontology#The_Clever_Fox_(Other_Animal)_1-69> ;
.
```

In this example, the reader can see how the type 6* is linked to a motif occurring typically in its storyline: we introduced for this a property called "linkToTMI". Additionally, the subclass relation is expressed, using the rdfs:subClassOf property. The "rdfs:label" property stores the original short title of the ATU type in English ("@en"). We encode the original description of the type as a value to the property "rdfs:isDefinedBy". A main aspect of the ontologisation of ATU (and TMI) is that each folktale type (or motif) is now represented by a Unique Resource Identifier (URI), and thus accessible in the Linked Data framework, once our data set has been published in its cloud.

An example of a motif ("K561.1") is given just below. We focused for the time being only on motif-ids and names. This current limitation is due to the inconsistent format of the motif descriptions and references used in the html code of the online resource, which made it difficult to be automatically extracted. We will include this information in a next version of the ontology. As pointed out earlier, elements of the TMI are encoded in a dual fashion: as belonging to the class "Motif" but also to its immediate non-terminal node (here "K561"). The rdfs annotation property "rdfs:label" is used for encoding the name of the motif (here in English, marked by "@en"). Multilingual correspondences can also be included as values of this property.



```
<http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-atu-ontology#K561.1>
  rdf:type :K561 ;
  rdf:type :Motif ;
  rdf:type owl:Class ;
  rdfs:comment "\"Index K561.1 of TMI\""@en ;
  rdfs:label "\"Animal captor persuaded to talk and release victim from his
mouth.\""@en ;
  rdfs:subClassOf :K561 ;
   :linkFromTMIToATU <http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-atu-
ontology#6* ;
```

In this example, we also see the property "linkFromTMIToATU", which is the inverse property of the one pointing between ATU elements and motifs. Additionally, we have introduced a third "linking" property, called "linkFromAaThToATU", which ensures that types of former versions of ATU are linked to the new names in the final version of ATU. By this final step of expanding our TMI-ATU ontology we ended up with the number of 14,937 classes and the number of 49,752 individuals that are interconnected by 3 object properties: "linkFromTMIToATU", "linkToTMI" and "linkFromAaThToATU". We managed thus to convert two valuable, handcrafted resources of literary knowledge consisting of more than 4000 pages into a 15.4 MB in size ontology file that can be easily accessed and searched.

Conclusions

We presented our work on the ontology generation for two widely used folktale classification systems. This ontology can be visualised and processed by standard OWL tools such as Protégé. The integrated ontology will be made openly available soon, after last quality controls. Current work is on adding as instances of the ontologies URLs of folktales that are marked with the corresponding numbers and so to allow access to those via Linked Data mechanisms.

Bibliography

Aarne, A. (1961). *The Types of the Folktale: A Classification and Bibliography.* The Finnish Academy of Science and Letters, Helsinki.

Ashliman, D. L. (1987). *A Guide to Folktales in the English Language: Based on the Aarne–Thompson Classification System.* New York, Greenwood Press.

Ofek, N., Darányi, S., and Rokach, L. (2013). *Linking Motif Sequences to Tale Types by Machine Learning.* Workshop on Computational Models of Narrative, 166-182.

Thompson, S. (1977). *The Folktale.* Berkeley: University of California Press.

Uther, H-J. (2004). *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson.* FF Communications no. 284–286. Helsinki: Suomalainen Tiedeakatemia.

Modeling Place in *Ulysses*: Ontologies and Pre-texts

Caleb Derven
caleb.derven@ul.ie
Maynooth University, Ireland

John Keating
john.keating@nuim.ie
Maynooth University, Ireland

Introduction

What role does place, if any, figure in the digital scholarly edition? Gunn and Hart offer a prescriptive view of the role of place in Joyce's *Ulysses*: "The topography of Dublin is "on the page" at least as much as are the meanings of the words "priest," "kidney," and ineluctable modality": it is a part of the book's primary reference system, without which its full sense cannot be apprehended" (Gunn et al. 2004). Such a positivistic view grants a key role to place and suggests that the novel may be modeled, to some extent, along geospatial lines. This paper proposes a model for foregrounding geographical elements, within a digital scholarly edition, for *Ulysses*.

However, rather than supporting the naturalistic view that *Ulysses* constitutes a sort of textual analogue to geographic representations of Dublin, previous work has suggested that *Ulysses* complicates this perspective (Bulson 2011). In this context Pierre Joris' formulation about the US poet Ronald Johnson, "He also knew that to make history you have to disfigure geography" might be applied equally to Joyce and the example of *Ulysses* (Joris 2009).

While significant work has been done to date outlining the extracting of geographical elements from texts and using the resulting toponyms in GIS contexts (Gregory and Hardie 2011, Cooper et al. 2015), this paper considers the question of toponyms in literary texts in both the context of modeling and the role of source texts in the representation of place. The key element of this research is making explicit the connections between source text and novel and how these connections are reflected in external ontologies.

Eide (2014) asserts that maps, texts and landscapes all constitute different ways of modelling and different ways of experiencing place. Texts have the characteristic of being "underspecified": e.g. a text might designate a place as "east" as opposed to exposing coordinates on a map. Maps and texts work to create the condition of intermediality - a combined geocommunication system. Are there source texts, relative to *Ulysses*, that amplify this sense of intermediality? Do these texts, in conjunction with the novel, occupy a sort of liminal category between text and map?

Methodology

This approach extends earlier work using TEI in effectively modelling geographical elements in one episode of the novel and incorporates alternative models such as ontologies and RDF XML documents in augmenting TEI XML (Derven et al. 2012). Such an approach serves to instantiate Gabler's claim for the digital scholarly edition as a "web of discourses" (Gabler 2010). The paper also considers whether this approach can be generalised beyond the specific use case of geographic named entities in the text and considers whether the combination of TEI XML, external ontologies and paratextual data stored in RDF triples may be more widely extended to digital scholarly editions. The approach taken in the paper extracts named entities from the text, identify toponyms and incorporates them as geo-rectified elements in an encoding of the novel. A sample of two episodes encoded in TEI XML is used. However, the approach to modelling taken extends beyond that offered by a strictly TEI-based approach by using a semantic web based approach and links placenames in the text to ontologies (for example, geoNames). The paper weighs marking up place names directly in the encoding itself against linking to external ontologies.

This paper also considers an interface for digital scholarly editions that accommodates multiple versions, displays geographic data extracted from the text and utilises an ontology to model and encode geographical elements within the text. The interface is used to interrogate a single episode from Joyce's *Ulysses*, a text that can be avowedly positioned within a geographic context. The tenth episode of the novel, Wandering Rocks, both foregrounds and privileges geographic elements in such a way that place itself takes on a narrative function. The use of place too alters slightly from manuscript to fair copy through to printed editions. The paper foregrounds the methodologies and tools used in assembling the interface. Toponyms are extracted, modelled in terms of both possible sources used in the construction of the text and existing ontologies, and presented as part of the interface for the digital scholarly edition through a GIS service. The interface includes a map that plots changes in the use of place name in the construction of the text.

Conclusions

In the context of a novel where place and place names function as narrative markers, investigating geographical elements in a computational context becomes another way to both read through the novel and assist in establishing a critical editing function. For example, considering the digital scholarly edition as interface, it becomes interesting to track such changes not only from the perspective of the critical edition but also in terms of GIS. How does the role and use of placenames change in the developing text? Is there evidence for paratextual or secondary sources in the development of place as this role changes? What is the most appropriate way to model and refer to place in the text? The paper, then, models episodes in *Ulysses* along these three dimensions: as a digital scholarly edition, as a mapping network, and through an ontology for modeling geographic elements.

Bibliography

Bulson, E. (2011) 'Disorienting Dublin' in *Making Space in the Works of James Joyce,* 1st ed, Routledge.

Cooper, D.C., Gregory, I.N., Hardie, A., Rayson, P. (2015) 'Spatializing and Analyzing Digital Texts: Corpora, GIS, and Places', available: http://e-space.mmu.ac.uk/579357/ [accessed 11 Aug 2016].

Derven, C., Teehan, A., Keating, J. (2015) 'Mapping and Unmapping Joyce: Geoparsing Wandering Rocks', Presented at the Digital Humanities 2014, Lausanne, Switzerland, available: http://dharchive.org/paper/DH2014/Paper-510.xml [accessed 20 July 2016].

Eide, Ø. (2014) 'Reading the Text, Walking the Terrain, Following the Map', in Arthur, P.L. and Bode, K., eds., *Advancing Digital Humanities,* Palgrave Macmillan UK, 194–205, available: http://link.springer.com/chapter/10.1057/9781137337016_13 [accessed 20 July 2016].

Gabler, H.W. (2010) 'Theorizing the Digital Scholarly Edition', *Literature Compass*, 7(2), 43–56.

Gregory, I.N., Hardie, A. (2011) 'Visual GISting: bringing together corpus linguistics and Geographical Information Systems', *Literary and Linguistic Computing,* 26(3), 297–314.

Gunn, I., Hart, C., Beck, H. (2004) *James Joyce's Dublin: A Topographical Guide to the Dublin of Ulysses : With 121 Illustrations,* Thames & Hudson, Limited.

Joris, P. (2009) *Justifying the Margins*, Salt Pub.

Joyce, J. (1986) *Ulysses: The Corrected Text*, reprint. ed, Vintage Books: New York.

Modernist Versions Project (2016) The Algorithmic Ulysses [online], *Modernist Versions Project,* available: http://web.uvic.ca/~mvp1922/the-algorithmic-ulysses/ [accessed 20 July 2016].

Travis, C. (2016) 'Bloomsday's Big Data', in *Literary Mapping in the Digital Age, Digital Research in the Arts and Humanities,* Routledge: London.

Automatically Analyzing Recordings of Musical Performances

Johanna Devaney
devaney.12@osu.edu
Ohio State University, United States of America

The Automatic Music Performance Analysis and Comparison Toolkit (AMPACT) is a suite of software tools for quantitatively analyzing musical performances for which a corresponding musical score is available. The primary target audience is music scholars who are interested in studying musical performance but who do not have the specialist technical skill or time required to develop their own tools. This toolkit enables music scholars to empirically, and automatically, analyze recorded performances, allowing scholars to study not just contemporary performance practices, but also the performance practices used in historical recordings. Empirical methods for analyzing musical scores have become increasing popular with the development of software such as Humdrum (Huron 1995) and music21 (Cuthbert and Ariza 2010), and the examination of recorded performance is a growing area of interest that shows similar potential (Cook 2014). AMPACT is particularly useful for musicologists and ethnomusicologists who are interested in undertaking longitudinal studies of music performances. Such studies may focus on large-scale differences in performances, such as the inclusion of additional notes, which is easy to detect by ear and annotate by hand. In contrast, they may focus the smaller scale, expressive aspects, of performances, such as variations in timing, dynamics, tuning, and timbre. The latter requires either careful, and extremely time consuming, manual analysis of visual representations of the audio sign or automatic analysis tools. Possible topics for smaller-scale studies include the development of a single performer's practice over time or changes in a particular performance style across time or geographical distance For general audiences in the humanities, the results of this type of longitudinal music performance research can be linked to other historical and/or geographic research in order to contextualize the similarities or differences observed across musical performances.

Researchers have studied recorded performances almost as long as recordable media has been available. Some of the most extensive work was undertaken by the psychologist Carl Seashore (1936, 1938) and colleagues at the University of Iowa also analyzed performances by pianists, violinists, and singers, employing a number of methods to study recorded performances. They studied piano performances from piano rolls and films of the hammers during the performance and violin and singing performances with a stroboscope, for frequency estimation, and an oscillograph, for intensity estimation. Accurate performance

data could be collected with these methods but the extraction of the data required a lot of work on the part of the researchers. As such, Seashore and his colleagues were only able to analyze a limited the number of pieces.

Interest in this type of detailed empirical analysis of recorded performances subsequently waned, likely due to the need for specialized equipment and the amount of time required for accurate analysis. Interest was revived in the late 1970s, as psychologists became intereste din studying musical rhythm, particularly in piano performance (see Palmer (1997) and Gabrielsson (1999, 2003)). The piano was popular because of its percussive nature and the ability to acquire measurements directly from the instrument—although this limited studies to performance made on specialized equipment. In contrast, the ability to extract performance data from recordings facilitates the study of existing recordings on a wider range of instruments.

There currently exist tools for accurately estimating tuning, and dynamics information for monophonic recordings once the locations of each note have been manually (e.g., PRAAT (Boersma 1993; Boersma 2001)) or automatically (e.g., Sonic Visualiser (Cannam et al. 2006) and TONY (Mauch et al. 2015)) annotated. For polyphonic audio (where there is more than one instrument playing), aligning the audio with a musical score can facilitate both the estimation of note locations and guide the subsequent estimation of tuning, dynamics, and timbre information for each note.

AMPACT is a suite of software tools uses score-guided methods estimate timing, tuning, dynamics, and timbre parameters for both monophonic and polyphonic audio. The technology implemented in AMPACT includes a score-audio alignment algorithm that is able to estimate timing asynchronies between notes which are written in simultaneities in the score (Devaney 2014), an important feature not only for studying differences in timing between performers but also for accurately guiding algorithms for estimating other performance parameters. The current version of AMPACT provides frame-level analysis of a number of tuning, dynamics, and timbre-related measurements for monophonic music and for tuning and dynamics for polyphonic music (Devaney and Mandel 2016). From these low-level measurements, high-level, perceptually informed, descriptors are calculated that summarize the low-level measurements over each note. Both the high-level descriptors and the low-level data can be stored in a special performance extension to the MEI encoding formats, so that the data can be released both for validating claims made in published research and for proving the data for other researchers to use.

In addition to describing AMPACT and its capabilities, this paper will also demonstrate the utility of AMPACT for digital musicology with a case study: a longitudinal study of the song "This land is your land" by Woodie Guthrie, a song that is notable for both its political and populist character. The case study explores systematic variations in the performance of the text "This land was made for you and me" within the performances (when the line retutns at the end of most verses) as well as across performers and performance contexts. Looking specifically at whether singers emphasize the words in different ways in live political contexts versus studio recordings. The performances include live and studio recordings by Woodie Guthrie, Pete Seeger, and Peter, Paul, and Mary, as well as more contemporary performers such as Wilco. AMPACT is used to automatically extract performance parameters (timing, dynamics, tuning) and compare the performance data within each recording and within each each performer's set of live and recordings to see how much variation occurs within songs, within singers' performances, and across performance contexts (e.g., studio, concert, and protest recordings).

Bibliography

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, 97–110.

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International* 5 (9/10): 341–5.

Cannam, C., Landone, C., Sandler, M.,and Bello, J. P. (2006). The Sonic Visualiser: A Visualisation Platform for Semantic Descriptors from Musical Signals*International Society of Music Information Retireval (ISMIR) Conference*.

Cook, N. (2014). *Beyond the Score: Music as Performance*: Oxford University Press.

Cuthbert, M. S., and Ariza, C. (2010). music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In *Proceedings of the International Symposium on Music Information Retrieval*, 637–42.

Devaney, J. (2014). Estimating onset and offset asynchronies in polyphonic audio-to-score alignment. *Journal of New Music Research* 43 (3): 266–75.

Devaney, J., and Mandel, M. (2016). Score-informed estimation of performance paramters form polyphonic audio using {A}{M}{P}{A}{C}{T}. In *Proceedings of the Late Breaking Demo Session of the International Society for Music Information Retrieval conference*

Gabrielsson, A. (1999). The performance of music. In *The Psychology of Music*, ed. D. Deutsch, 501–602. Original edition, San Diego, CA: Academic Press.

Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of Music* 31(3): 221–72.

Huron, D. (1995). *The Humdrum Toolkit: Reference Manual*. Menlo Park, California: Center for Computer Assisted Research in the Humanities.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., and Dixon, S. (2015). Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proceedings of the Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, 23-30.

Palmer, C. (1997). Music performance. *Annual Review of Psychology* 48 115–38.

Seashore, C. (1936). *Objective Analysis of Musical Performance*. Iowa City, IA: University of Iowa Press.

Seashore, C. (1938). *Psychology of Music*. Iowa City, IA: University of Iowa Press. Original edition, New York, NY: Dover Publications.

Leveraging Expert Domain Knowledge to Learn a Representation of Symbolic Music

Johanna Devaney
devaney.12@osu.edu
Ohio State University, United States of America

Music manuscripts offer as much potential as text manuscripts for data mining and, as with resources like Google Books, there is a wealth of data available online. Currently the largest resource, the International Music Score Library Project (IMSLP), has more than 370,000 scores in its database. While many of these scores exist only in image-based formats, ongoing improvements in the area of Optical Music Recognition (OMR) allow for automatic conversion from images to symbolic representations, which include information such as instrumentation, key signature, time signature, and the notes' pitches, metrical positions, and durations. In order to exploit the research potential of these symbolic music databases, a representation that captures temporal relationships within the music is needed that highlights the structurally significant parts of the musical surface, while ignoring ornamentation.

This paper describes a representation that emphasizes the more structurally significant parts of the musical surface and de-emphasizes less significant parts, such as ornamentation, by integrating human domain expertise and data-driven approaches within a temporal machine learning model. The representation contains less information than the musical surface but more than corresponding chord labels, which discard information about musical texture and are too generalized to use for detailed similarity and classification tasks. The weighting for the various components of the musical surface is determined from an initial harmonic analysis. This harmonic analysis will be performed by a hierarchical model of chord labels and phrases, which will function like a "language model" in speech recognition. In music theory, phrase models describe musical phrases in terms of the tonic, predominant, and dominant functions (Laitz 2012). The inclusion of the expert domain knowledge expressed in the phrase function model helps to resolve the ambiguity between the musical surface and appropriate chord labels in the harmonic analysis, namely whether a particular chord is likely to occur in a particular part of a phrase. Taken in combination with OMR, this representation could be used to render searchable all available scanned music. These searches would not be limited to melody, as is the current state-of-the-art, but would also allow for querying by chord progressions and/or formal structures. The representation can also facilitate automatic hierarchical analysis of musical structures and provides a basis from which to undertake classifications and similarity tasks. Classification tasks include harmonic analysis or assessing the likelihood of a particular composer having composed a piece of unknown provenance, while similarity tasks include longitudinal studies over a composer's career or across composers.

Much of the work on analyzing the growing wealth of music data has been heavily influenced by text retrieval methods through their use of N-grams, sequences of N contiguous symbols. N-grams work well in modeling monophonic sequences, such as when directly applied to the musical surface for monophonic melody retrieval (Pickens, 2001) and for chord retrieval when the chords occur as distinct vertical units (Scholz et al., 2009). This has been demonstrated effectively on peachnote.com (Viro, 2011) with an N-gram viewer similar to the one Google makes available for Google Books. One significant area where N-grams have problems, however, is for more complex textures where the notes of chords are not played simultaneously, which is true of a large proportion of western art music since 1750. One way to address this problem is to automatically segment the musical surface into beat-length frames and treat the contents of each frame as a "chord" (Radicioni and Espositio, 2006), which is well suited to chorale textures but is problematic for arpeggiations or other textures where the chords notes don't occur simultaneously. Another approach is to analyze chord labels rather than the musical surface, such as the system of de Haas et. al (2011), although these are often not available.

The representation described in this paper takes a different approach, using a conditional random fields (CRFs) model for developing both a data-driven model, where all of the feature functions and potentials are learned from the data, and a hybrid data-driven/rule- driven approach, where domain knowledge "rules" are used to design feature and potential functions. Data for the purely data-driven approach comes from a domain expert-labeled dataset of Mozart and Beethoven piano works in theme and variation form (Devaney et al. 2015). The rule-driven approach incorporates the rules presented in textbooks used in undergraduate music theory curricula, primarily Laitz (2012). This hybrid data- and rule- driven approach is motivated by previous work that demonstrated that a combination of data- and rule-driven models performed better than either approach alone on music analysis tasks (Devaney and Shanahan 2013).

This paper will also discuss the implications of this use CRFs for analyzing other metrically structured cultural products, such as poetry or song lyrics, as well as how this approach could be generalized to other digital humanities projects, specifically for relatively "data- poor" problems where there is a large amount of domain expertise that can be modeled, such as the study of narrative in natural language. More broadly, this work presents a vision of the digital humanities, where large-scale data-driven approaches are balanced by deep domain knowledge and the types of humanistic questions being asked require the development of more sophisticated technology than is currently available.

Bibliography

de Haas, W.B., J.P. Magalhaes, R.C. Veltkamp, F. Wiering. (2011). HarmTrace: Improving Harmonic similarity estimation using functional harmony analysis. In *Proceedings of International Society of Music Information Retrieval conference (ISMIR)*. 67–72.

Devaney, J., C. Arthur, N. Condit-Schultz, and K. Nisula. (2015). Theme And Variation Encodings with Roman Numerals (TAVERN): A new data set for symbolic music analysis. In *Proceedings of the International Society for Music Information Retrieval (ISMIR) conference*, 728–34.

Devaney, J., and D. Shanahan. (2014). Evaluating Rule- and Exemplar-Based Computational Approaches to Modeling Harmonic Function in Music Theory Pedagogy. In *Proceedings of the 9th Conference on Interdisciplinary Musicology.*

Laitz, S. G. (2011). *The Complete Musician*. Oxford: Oxford University Press, 3rd edition. Pickens, J. 2001. A survey of feature selection techniques for music information retrieval.

Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts, Technical Report.

Radicioni, D.P. and Esposito, R. (2006). Learning tonal harmony from Bach chorales. In *Proceedings of the International Conference on Cognitive Modelling*.

Scholz, R., Vincent, E., and F. Bimbot, F. (2009). Robust modeling of musical chord sequences using probabilistic N-grams. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 53–6.

Viro, V. (2011). Peachnote: Music score search and analysis platform. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. 359–62.

L'extraction automatique des motifs dramaturgiques dans les séquences de deux scènes

Marc Douguet
marc.douguet@paris-sorbonne.fr
OBVIL, Université de Paris-Sorbonne, France

La structuration intrinsèque du texte théâtral en une suite de scènes réunissant chacune un nombre défini de personnages offre des potentialités infinies pour l'analyse quantitative. Ce domaine a déjà fait l'objet de nombreuses études, notamment dans la lignée des travaux de S. Marcus, qui montre qu'une pièce de théâtre peut être modélisée sous la forme d'une matrice binaire où chaque personnage est codé comme une suite de 0 et 1 selon qu'il est absent ou présent dans les scènes successives qui divisent le texte (Marcus 1973).

Jusqu'à présent, l'analyse des matrices dramatiques s'est notamment concentrée sur le nombre de scènes où apparaît un personnage, son nombre d'entrées, les différentes relations de coprésence qu'il entretient avec d'autres personnages, etc. (Marcus 1972; Lafon 1990; Ilsemann 1997) Ces paramètres permettent de qualifier de manière globale les personnages et l'intrigue en appliquant à la dramaturgie des outils empruntés à l'analyse statistique ou à l'analyse des réseaux. Cependant, selon cette approche, l'enchaînement des scènes et l'ordre dans lequel elles sont disposées ne sont pas pris en compte. Ils le sont en revanche dans les travaux de M. Dinu sur la stratégie des personnages, qui cherchent à calculer la probabilité de la réalisation d'une configuration de personnages donnée en fonction de celles qui la précèdent (Dinu 1984).

Nous voudrions ici présenter une nouvelle approche qui, comme celle M. Dinu, prend en compte la progression du texte, mais adopte le point de vue surplombant qui est celui du dramaturge au moment où il dresse le plan de sa pièce et s'intéresse avant tout aux techniques d'écriture dramatique. Nous introduisons pour cela la notion de *motif dramaturgique*, défini comme une séquence d'entrées et de sorties de personnages, abstraction faite de l'identité de ces derniers. Deux suite de scènes ne faisant intervenir aucun personnage commun ou appartenant à des pièces différentes peuvent donc instancier le même motif.

Prenons comme exemple la matrice dramatique de *Mélite* de Corneille :

Actes	I					II								III						IV										V					
Scènes	1	2	3	4	5	1	2	3	4	5	6	7	8	1	2	3	4	5	6	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6
TIRCIS	1	1	1	0	1	0	0	0	1	0	0	1	1	**0**	**1**	**1**	**1**	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
ERASTE	1	1	1	0	0	1	1	1	0	1	1	1	0	**0**	**0**	**0**	**0**	**0**	0	0	0	0	0	1	1	**0**	**1**	**1**	0	1	1	0	0	0	1
MÉLITE	0	1	0	0	0	0	1	0	0	0	0	1	1	**0**	**0**	**0**	**0**	**0**	0	1	1	1	1	0	0	**0**	**0**	**0**	0	0	0	0	1	1	1
CHLORIS	0	0	0	1	1	0	0	0	1	0	0	0	0	**0**	**0**	**0**	**1**	**1**	1	0	1	1	1	0	0	**0**	**0**	**0**	1	0	0	1	0	1	1
PHILANDRE	0	0	0	1	1	0	0	0	0	0	1	0	0	**1**	**1**	**0**	0	0	1	0	0	0	0	0	0	**1**	**1**	**0**	0	0	0	1	0	0	0
CLITON	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
LA NOURRICE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	1
LISIS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0

Les scènes III, 1-3, III, 3-5 et IV, 7-9 instancient un même motif, représenté en gras, qui peut être exprimé sous la forme d'une sous-matrice

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

ou sous la forme plus lisible d'une séquence codifiée de caractères où A désigne le personnage qui apparaît en premier, et B celui qui apparaît en second, les scènes étant délimitées par "/" et les personnages, à l'intérieur d'une scène, par "-" :

A/A-B/B

La comparaison de ce motif avec un autre motif de trois scènes et deux personnages permet de comprendre l'intérêt de cette méthode. Ce motif, que l'on rencontre dans *Mélite*, II, 1-3, est le suivant :

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

soit

A/A-B/A

Notre hypothèse est qu'indépendamment de la longueur de chaque scène composant le motif et de la répartition de la parole entre les personnages, chaque motif présente toujours, dans la diversité de ses réalisations, des enjeux comparables. En l'occurrence, les deux motifs que nous venons de citer illustrent deux techniques de composition bien distinctes. Soit un personnage reste en permanence sur le plateau, soit le dramaturge renouvelle entièrement le personnel dramatique, profitant de la scène centrale pour faire se croiser les personnages qui se relaient sous les yeux des spectateurs. Dans le premier cas, un personnage sert de pivot à l'action, et en rencontre successivement plusieurs autres. Ce faisant, il guide l'attention du public. Dans le second, le point de vue du spectateur dépasse celui de l'ensemble des personnages : il en sait et en voit plus que chacun d'eux individuellement.

L'étude des motifs dramaturgiques devient encore plus intéressante quand on regroupe ensemble les personnages qui, à l'intérieur d'un motif, ont une distance scénique nulle

(Marcus, 1972), c'est-à-dire sont toujours présents et absents en même temps. Ainsi, *Mélite*, I, 1-3 repose sur le motif

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Or malgré le fait qu'il y a deux personnages présents en permanence, et non un seul, ce motif n'est pas fondamentalement différent de

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Dans les deux cas, une confrontation centrale est entourée par deux scènes présentant un caractère plus "privé", qui permettent de comparer l'attitude d'un personnage ou groupe de personnages avant et après une rencontre avec un autre personnage ou groupe de personnages et de faire entendre le discours qu'ils tiennent en l'absence (et bien souvent au sujet) de ce dernier.

Méthodologiquement, cette approche est comparable à l'étude des motifs syntaxiques développée dans le cadre de la stylistique computationnelle (Boukhaled 2015), si ce n'est que chaque élément du motif se définit non par le choix unique d'une classe de mot, mais par le choix multiple d'un ensemble de personnages. Elle possède donc *a priori* les mêmes applications, notamment la qualification du style d'un auteur (entendu ici comme style de composition, et non comme style d'écriture, et dont l'unité de base n'est pas le mot, mais la scène).

Nous avons constitué une base de données comportant tous les motifs de moins de huit scènes extraits d'un corpus de plus de 200 pièces de théâtre françaises écrites entre 1630 et 1680. Le choix de cette période se justifie par son importance dans l'histoire du théâtre français, et par le renouvellement des techniques dramaturgiques qu'elle permet d'étudier (passage du théâtre "baroque" au théâtre "classique"). Pour interroger cette base, nous avons développé un moteur de recherche qui, pour un motif donné, renvoie sa fréquence d'apparition sur l'ensemble du corpus, son évolution, ainsi que sa répartition par genre et par auteur.

Il ne s'agira pas ici de présenter les détails techniques de ce travail, mais de livrer les premières conclusions auxquelles nous sommes parvenus. Une des pistes possibles pour exploiter cette base de données est d'étudier l'évolution et la répartition de l'ensemble des motifs possibles pour un nombre donné de personnages et un nombre donné de scènes. Pour cette première étude, nous nous intéresserons aux motifs composés de deux scènes. La combinatoire est facile à produire. Si l'on regroupe les personnages qui sont présents et absents en même temps, seuls trois personnages ou groupes de personnages peuvent intervenir (un quatrième personnage aurait nécessairement une distance scénique nulle avec un autre), et quatre motifs sont possibles:

A/A-B
A-B/A
A/B
A-B/A-C

On constate, dans notre corpus, la disparition progressive du motif A/B (aucun personnage commun entre deux scènes contiguës), qui s'explique par l'exigence croissante de continuité dramatique. Un autre phénomène est plus inattendu et n'a pas encore, à notre connaissance, été vraiment étudié : deux scènes sont plus souvent reliées par une entrée (A/A-B) que par une sortie (A-B/A). Autrement dit, les personnages ont tendance à entrer séparément, mais à sortir de manière groupée, ce que nous tenterons d'expliquer en convoquant la notion de "tension dramatique". Enfin, on constate une légère augmentation puis une légère baisse des entrées et des sorties simultanées (A-B/A-C), dont nous essaierons de rendre compte en mettant ce phénomène en corrélation avec l'apparition d'une technique d'enchaînement plus complexe reposant sur un tuilage des entrées et des sorties à l'échelle de trois scènes successives : pour relier deux dialogues différents, les dramaturges préfèrent de plus en plus intercaler une scène de transition à trois personnages (A-B/A-B-C/A-C) plutôt que faire entrer et sortir deux personnages en même temps sans que ceux-ci ne se parlent (A-B/A-C).

Ces considérations générales nous mèneront à étudier le détail de la répartition des quatre motifs par genre et par auteur. Enfin, nous distinguerons les scènes qui appartiennent au même acte et celles qui se situent de part et d'autre d'un entracte. Les entractes sont en effet rarement pris en compte dans l'analyse des matrices dramatiques, alors qu'ils jouent un rôle essentiel. Une des règles de la dramaturgie classique voudrait par exemple qu'un personnage présent à la fin d'un acte ne le soit pas au début du suivant (Aubignac 1657). Une étude statistique vérifie en partie cette règle, mais montre aussi qu'elle est très fréquemment enfreinte.

Toutes ces questions visent à définir une norme d'écriture (pour une époque, un genre, un auteur). Mais les outils numériques ne se limitent pas à l'analyse quantitative : ils enrichissent également l'analyse littéraire en permettant de découvrir à l'intérieur d'un vaste corpus des cas particuliers, qu'il serait quasiment impossible de repérer sans procéder à une extraction automatique. On se penchera ainsi sur les pièces où l'un des quatre motifs est sur- ou sous-représenté, en se demandant quels sont les enjeux de cette anomalie pour la conduite de l'action et l'effet produit sur le spectateur.

Bibliographie

Aubignac, abbé d' (1657). *La Pratique du théâtre*. Paris: Sommaville.

Boukhaled, A., Frontini, F. et Ganascia, J.-G. (2015). Une Mesure d'intérêt à base de surreprésentation pour l'extraction des motifs syntaxiques stylistiques. *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*.

Dinu, M. (1984). Entropy and Prediction in the Study of Theatre.

Poetics 13, pp. 57-70.

Ilsemann, H. (1995). Computerized Drama Analysis. *Literary and Linguistic Computing* 10(1), pp. 11-21.

Lafon, D. (1990). *Le Chiffre scénique dans la dramaturgie moliéresque*. Ottawa/Paris: Presses de l'Université d'Ottawa/Klincksieck.

Marcus, S. (1973). *Mathematische Poetik*. Bucureşti/Frankfurt am Main: Editura Academiei/Athenäum Verlag.

Marcus, S. (1972). Stratégie des personnages dramatiques. *Sémiologie de la représentation*. Bruxelles: Complexe, pp. 73-95.

"Come Together, Right Now": Discovery and Interoperability for Born-Digital Music Scholarship

Timothy Charles Duguid
tim.duguid@glasgow.ac.uk
University of Glasgow, United Kingdom

Elizabeth Grumbach
egrumbac@tamu.edu
Texas A&M University, United States of America

Summary

Music scholars have unprecedented access to an ever-growing digital corpus of music-related content, as libraries and institutions continue to digitize their holdings at an extraordinary pace. Yet, access to and discovery of these resources is problematic for music scholars. In 2016, the Music Scholarship Online (MuSO) project was awarded a grant from the National Endowment for the Humanities, Office of Digital Humanities to begin forming a community of music scholars to address these issues. This short paper will present the results of the planning meeting and further work done to address the social and technical issues impeding the interoperability, discovery, and access of born-digital music scholarship.

Proposal

Music scholars have unprecedented access to an ever-growing digital corpus of music-related content, as libraries and institutions continue to digitize their holdings at an extraordinary pace. Of course, curators describe digital content in ways that best suit their collection's needs, using a variety of metadata standards such as MARC, Dublin Core, and LIDO (just to name a few). Despite the obvious benefits to such flexibility, interoperability between digital collections has proven difficult, forcing music scholars to navigate between several interfaces and platforms in order to locate research materials. More importantly, access to these resources is limited to scholars already aware of their existence.

Platforms such as the Digital Public Library of America (DPLA), Europeana, and HathiTrust are working to solve these issues by aggregating digital collections into single catalogues, therefore making collection holdings accessible and interoperable. These platforms and others like them are adept at aggregating a variety and breadth of content (including, for instance, audio recordings, encoded content, musical scores, photographs, manuscripts, and printed texts) relevant to a variety of disciplines. The structure of these aggregators are heavily reflective of the library collections that make up the bulk of their data. But, unlike the discrete digital collections themselves, these platforms bring together information about holdings and digitized materials from diverse institutions, borders, and economies in order to make them discoverable in new ways.

However, there is a host of content consistently left out of even the best and most extensive music catalogues and digital aggregators: born-digital scholarship. This new model for of scholarly knowledge production has motivated humanities scholars to consider how digital catalogs and collections should evaluate and make visible new forms of scholarship. When Jerome McGann introduces the DPLA in 2011's "On Creating a Usable Future," he locates our humanities crisis not only in the evaluation of digital scholarship, but also "the sustainability of born digital resources and the work they support" and forging paths to make new forms of scholarly knowledge production "a general institutional practice" (182). To make digital scholarship an institutional practice, we must consider not only our social infrastructures (promotion and tenure), but the technical infrastructures that allow discoverability and interoperability - two concepts that are crucial to making our work visible and accessible across audiences and publics. For digital scholarship in music, no aggregators, platforms, or institutional ventures currently ensure that born-digital scholarship is widely discoverable and therefore accessible.

Music Scholarship Online (MuSO) is a community of scholars dedicated to resolving these issues, thus providing greater access to born digital music scholarship. It began in 2015 as a Digital Humanities Start-up Grant funded by the National Endowment for the Humanities, in which music librarians, music encoders, and musicologists gathered to discuss issues surrounding aggregation and peer review for born-digital music scholarship. Participants in the planning meeting came together to ask and answer "What do music scholars need in a digital curator and search mechanism?" Ultimately the team chose to follow the model of the Networked Infrastructure for Nineteenth Century Electronic Scholarship (NINES) and the other members of Advanced Research Consortium (ARC). In 2016, MuSO officially joined ARC, a federation of virtual research environments that coordinates several thematic and period-specific aggregators, like MuSO, into a catalog "containing resources spanning the bulk of Western civilization, from the medieval period to the early 20th-century" (Grumbach and Mandell, 3).

This short paper describes the present metadata prototyping efforts of the Music Scholarship Online (MuSO) project. Arguing that it is unnecessary for digital project teams to generate preservation-quality metadata, this paper describes how MuSO is working with ARC to generate a metadata schema for discovery that is lightweight and therefore a better option for digital projects facing constrained budgets and limited timescales.

Beginning with a brief introduction to the unique challenges facing born-digital projects in music, the paper then presents the history of MuSO and its initial phase, in which the basic metadata guidelines and peer review processes were developed. Afterwards, the paper reports on current efforts to build a prototype schema for MuSO: a discovery-level metadata standard for digital projects in music. It shows how the MuSO and ARC leadership teams have worked together in examining present aggregators of music resources such as the the Digital Archive of the Beethoven-Haus, the Digital Image Archive for Medieval Music, and the Juilliard Manuscript Collection, as well as well-established catalogues such as *Répertoire Internationale des Sources Musicales* (RISM) and *Répertoire Internationale de Littérature Musicale* (RILM). It reveals the similarities and differences between the metadata standards of these resources, and then it identifies the most significant elements for discovery-level metadata.The paper then concludes by comparing these elements with those present in the musical holdings of multidisciplinary collections such as the DPLA, Europeana, and the HathiTrust to ensure that
MuSO's lightweight metadata schema accurately captures the level of description needed to work within the multidisciplinary ARC catalog.

If born-digital scholarship is to become general institutional practice, the methods of discovery for that scholarship must be familiar to all humanists: all of the multidisciplinary, multilingual and international scholars that our institutions serve. The authors will therefore present findings concluded from an examination of the current MuSO Metadata recommendations as decided by the MuSO community alongside existing multidisciplinary metadata schemas. By investigating the push and pull between discovery-level metadata and preservation-level metadata, this short paper reports our efforts to ensure that Music Scholarship Online effectively describes the born-digital scholarship it intends to aggregate.

Bibliography

Grumbach, E., & Mandell, L. (2014) "Meeting Scholars Where They Are: The Advanced Research Consortium (ARC) and a Social Humanities Infrastructure."*Scholarly and Research Communication* [Online], 5.4: n. pag. Web. 1 Nov. 2016

McGann, J. (2011) "On Creating a Usable Future." *Profession:* 182-195. New York, NY: Modern Language Association of America.

How the notion of access guides the organization of a European research infrastructure: the example of DARIAH

Suzanne Dumouchel
suzanne.dumouchel@dariah.eu
DARIAH, France

DARIAH (Digital Research Infrastructure for the Arts and Humanities) is, as the name implies, an infrastructure dedicated to research in digital arts and humanities. Developed under the auspices of the European Commission, it aims to organize communities in those fields, to develop interdisciplinary projects, promoting in particular the digital dimension of humanities and arts research by disseminating good practices, and providing tools and services. In legal terms, it is what is called an ERIC, that is to say a European Research Infrastructure Consortium, which is composed of national members that have come together to promote common objectives and serve common communities. Several ERICs have been created since 2009 in Europe, mostly based on a disciplinary approach, membership by several European countries, and the pooling of services.

DARIAH had already a long existence as an unofficial structure since 2005, but it become fully established as an ERIC in 2014. It brings together 17 countries in Europe which makes it the biggest ERIC in terms of members, but it is also distinct because it serves a very wide community consisting of the whole of Arts and Humanities research. This breadth poses real questions regarding the notion of access. The role of all ERICs is to share tools. services, human resources, projects, software, etc. to enrich research. Doing so for a community as broad as DARIAH's creates particular challenges.

Through this presentation, we wish to give an account of the specificities of this infrastructure by presenting the issues related to the notion of access that contribute to the structuring of DARIAH. Many of our pre-existing understandings of how access impacts upon humanities and arts research date from the days of the library collection, and it is common to limit the notion of access to research data. But within the framework of a research infrastructure consortium, the notion of access is made more complex. We will evidence this different paradigm here with five examples, which we will develop in turn by explaining both the constraints and the solutions that are envisaged to solve the problems encountered.

According to the Cambridge Dictionary, "access" has two kinds of meanings: the first one concerns "the method or possibility of getting near a place or a person” and the second one “the right or opportunity to use or look at something”. Both meanings are interesting in terms of creating an expanded understanding of access in field of Digital Humanities. Even if we mostly think about the second one (open access, open data, and so on), the first one shows the necessity of being near our community and highlights the fact that thanks to digital tools we are nearer and nearer despite the distance and we are able to work together being in different places.

DARIAH positions itself to support an emerging research culture in ways that invoke both of these meanings, as the examples below illustrate.

Managing interdisciplinarity

When one addresses a community as broad as that encompassed by the expression "Arts and Humanities", itself vague regarding the disciplines that it covers, one implicitly raises the question of what access by these Communities to this hyper-infrastructure represented by DARIAH will mean.

This implies first and foremost the need to define, more or less precisely, what is meant by the expression "Arts and Humanities", and in particular the expression "Humanities", which varies across European languages and across places and times. For example, are the social sciences included or not? And on what grounds can you bring together researchers from communities as diverse as literature, history, philosophy, cinema studies and perhaps even geography and linguistics? To be useful, DARIAH needs to develop a common language with common services and tools which can be used by people in those different fields. In this case, access concerns the way DARIAH communicates with its communities and the projects it launches, to encourage interdisciplinarity without seeming intrusive. Indeed, this infrastructure has as one of its objectives to contribute to organizing a gigantic network of specialists in those fields. To do that, it is necessary to think about how to create such a network, to animate it and to make it last. Several initiatives are being developed, both thematic and national, which we will present as responses to this challenge.

Managing tensions between national and international perspectives

Access also has a political dimension. When the same tool is developed in parallel in two different countries, how can we know how to assign credit for the development? Which should be valued by DARIAH? There is thus a problem of selection, and therefore a possibility of bias, in the choice to favor promoting the access to one tool rather than another. To resolve this tension, the DARIAH community is coordinated around National Representatives who are engaged in complementary, rather than competitive, work. This work is based in particular on the dissemination of information about the activities of the national teams and the projects in which they are involved.

On the other hand, some teams may wish to retain the rights or control of their tools and may not wish to make them accessible to other communities without compensation. We will see how this constraint contributes in turn to the structuring of the DARIAH ERIC.

The question of access thus raises political and diplomatic problems that may interfere with more neutral criteria of quality of the tool, its durability, its usefulness.

Speaking to whom?

The role of DARIAH is also very broad insofar as this European consortium has an ancillary mission for the development of new communities. But this mandate is very vague. Do we mean fields of research? Specific countries? Or people inside? In this section, we would like to focus on people. Toward this end, DARIAH has specific functions in terms of teaching digital practices, not only within the current network but also beyond it, with the goal of opening and expanding it to encompass researchers (including under and postgraduate students) who may or may not have any competences in digital tools but who are interested in them. In this way DARIAH acts as a facilitator to help people to use digital tools and services.

One question that we must ask, however, is whether we can or should open our community to, for example, private companies, which would help in the development of tools and / or which would benefit, once again, from inclusion. The wider the access, the less control there is. And what would DARIAH mean and how would it act if the infrastructure became open to all without distinction of disciplines, places, people? Conversely, what would be its meaning, if by privileging shared access to knowledge and tools, it decided to close the door to some? Imperatives toward democratizing the benefits DARIAH can bring come at such junctures into conflict with the possibility that too much access could dilute the infrastructure’s effectiveness, distort its scale or divert its mission. Again, this is an aspect related to the question of access to which the infrastructure must respond and on which we shall give a few quick lines of reflection.

Managing tools

This point is particularly well recognised within the DH community, since it questions the interoperability of tools. For DARIAH, questions regarding managing access to tools arises in terms of languages, content, formats and, of course, sustainability. Within DARIAH, the issue of interoperability is paramount, to leverage our large scale, but also to enable disciplinary practices usage models; given that the research questions posed at the origin of these uses will vary so considerably.

Specific attention is therefore paid to this aspect and in particular to data hosting. One of the first tasks that DARIAH has set itself is to work on long-term data hosting. To do this, it has, for example, relied on national hosts able to also integrate data from multiple countries. These include the CNR in Italy (via the PARTHENOS project) and Huma-Num in France.

This perspective on access reflects as well the importance of the trust that must be established between the partner countries, in particular with regard to intellectual property, as suggested in the next and final point.

Building collaborative tools

DARIAH is an infrastructure that brings together 17 countries and a range of diverse disciplinary communities. In this sense, it involves collaborative work that relies mainly on the use of digital tools. But one problem remains: the too easy access to collaborative tools developed by companies that do not share the same conception of intellectual property and data security. Tools such as Google doc and Google drive, etc. are unavoidable in the context of collaboration between researchers, but the access in this case is so easy that their existence prevents the development of alternative tools that correspond more closely to the specificities of scientific exchanges. It is now important to develop virtual working environments conducive to scientific exchanges and the needs of researchers, particularly in the communities concerned.

To enhance its ability to navigate the many requirements of access, DARIAH has recently launched a far-reaching 3-year program of actions, which will be presented as well as an example of how a holistic approach to access can be manifested in an institutional strategy. By explaining how the document has been formulated and how community support for it has been developed, the presentation will give a worked example of how access can be negotiated across countries and disciplines.

The notion of access lies at the heart of the issues dealt with by infrastructures such as DARIAH, as they seek to structure and facilitate coordination and exchange of tools, and the development of research. Questions of access, which are too often reduced to the management of the data, imply each time a positioning; And that even when the stated objective is to be open to all, it is nonetheless subject to a form of choice.

Bibliography

Tibor Kalman, T., Wandl-Vogt, E. (2014) DARIAH-ERIC Towards a sustainable social and technical European eResearch Infrastructure for the Arts and Humanities. *e-IRG Workshop*, Nov 2014, Rome, Italy. <http://www.e-irg.eu/e-irg-events/workshop-10-11-november.html>. <hal-01081479>

Oltersdorf, J., Matoni, M., Thiel, C. (2016) DARIAH Report on researchers' service needs. [Other] DARIAH. <hal-01351267>

Romary, L. (2011) Partnerships, relationships and associated initiatives — Towards a strategic plan for DARIAH. [Research Report] R EU 4.3.1, DARIAH. <hal-01150112>

Romary, L., Mertens, M., Baillot, A. (2016). Data fluidity in DARIAH – pushing the agenda forward. *BIBLIOTHEK Forschung und Praxis*, De Gruyter. 39 (3), pp.350-357. <hal-01285917v2>

Romary, L., Chambers, S. (2014) DARIAH: Advancing a digital revolution in the arts and humanities across Europe. *e-data&research*, Data Archiving and Networked Services (DANS), 2014. <hal-00913691>

Augmented dance scholarship: computer-assisted analysis of Javanese dance

Miguel Escobar Varela
m.escobar@nus.edu.sg
National University of Singapore, Singapore

Luis Carlos Hernández Barraza
a0107963@u.nus.edu
National University of Singapore, Singapore

This case study of Javanese dance argues that biomechanical analysis of dance can be used to augment dance scholarship: quantitative biomechanical data can complement rather than supplant other kinds of dance scholarship. In proposing this, we draw a parallel to the vision of computer assisted-literary criticism presented by Sinclair and Rockwell (2016), where interpretive textual analytics enable different readings of a text or corpus, rather than a quest for scientific truth. Similarly, we suggest that the visualization of biomechanical data enables other ways of looking at dance.

Our research project continues a long history of recording dance movements, but also departs from previous projects in important ways. Different notation systems, recording media and technological tools have been developed in order to capture the fleeting quality of dances. A few well-known examples include Eadweard Muybridge's photographic series depicting dance movements in the 19th century, the Labanotation system developed by Rudolf von Laban in the early 20th century and the application of biomechanics to study ballet in the 1970s, as exemplified by the papers in the conference *At the Dance: Verities, Values, Visions: Binational Dance Conference* in Waterloo, Ontario in 1971.

The practical applications of these systems and methods have been justified along several lines: the transmission of choreographies, dance pedagogy, description and classification of dance movements, and injury prevention. Initial enthusiasm for the scientific study of dance waned after the 1970s and has had little impact in mainstream dance scholarship. Recording of choreographies using Labanotation and motion-capture devices are still very common, but the quantitative analysis of the data obtained in this way is not.

In a parallel development, choreographers and dance collectives often use dance-tracking systems and visualizations as part of their work; notable examples of such practitioners include William Forsythe, Siohban Davis, Merce Cunningham, Anne Teresa De Keersmaeker and BADco. Unlike the quantitative studies of dance, these creative uses of dance data within performances are widely celebrated within dance scholarship and practice (Bleeker 2016). These playful uses of data and visualization are mostly constrained to creative dance practice. Even if they are often discussed within dance scholarship they are not often deployed as analytical approaches within dance scholarship. Our present study shows a way in which this could change.

Combining methods from dance ethnography and biomechanical engineering, we recorded differences in the building blocks of a well-known Javanese dance: *Sendratari Ramayana*. The *Sendratari* (often translated as ballet) is a modern Indonesian dance practice that combines movement gestures from royal dance traditions (*beksa* and *srimpi*), and from classical operatic dance-dramas (*wayang wong*) in order to present stories from the Ramayana (the Sanskrit epic that provides narrative background for many performance forms in South and Southeast Asia). We were interested in the "body types", a series of rules that describe how the main five types of characters should move: *luruh* (extremely refined), *lanyap* (vigorous-refined), *gagah* (vigorous), *raksasa* (ogres), *wanara* (simians), and *manuk* (birds). These rules define the rhythm and body angles at which the same movements (standing, walking, fighting, etc.) are performed by each character.

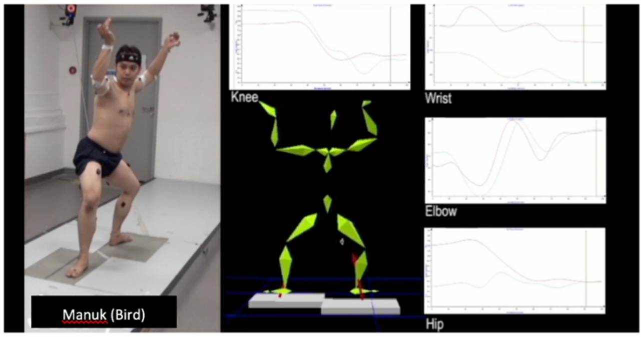


Figure 1. The dancer performs the standing motion corresponding to a bird. The movement can be seen as an animation of all major limbs and as graphs carting the movement of each joint through time.

For our study, we placed markers in all major joints of a professional dancer and we asked him to perform all basic movements as would befit each body type. Using the resulting data, we calculated the range of motion for each joint (difference between the maximum and the lowest angles). By comparing the ranges of motion pertaining to different body types, we discovered that the more subtle, elegant body types have the biggest range of motion. This is surprising even to experienced dance scholars. The demons and monkeys hold their arms in wider gestures and it would be easy to conclude that they have wider ranges of motion. The movement of the refined characters is slower and one can easily lose track of the fact that these movements actually require a greater range of motion. Indeed, similar assertions are absent in the literature. Our goal is to present our results through two interactive web displays:

- As a series of video panels, where a user can see a particular movement performed by a particular body type. Different videos can be displayed together: a recording of the dancer, an animation showing the position of the markers, and animated graphs that show the movement of specific joints (Fig. 1).
- Graphs that show a comparative overview of all movement types. For example, using Principal Component Analysis (a method often used in textual scholarship in the DH), we generated a comparison of the different body types (Fig. 2).

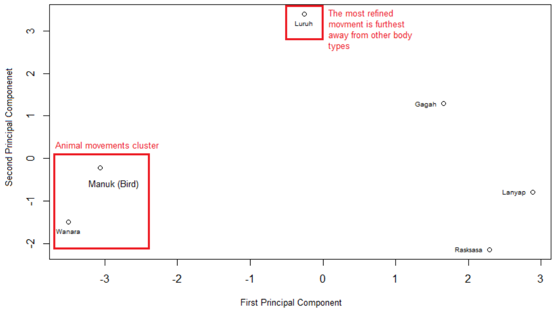


Figure 2. Principal Component Analysis of the ranges of motion of the main body types.

These results are not only aimed at a quantitative validation of scientific premises: their key contribution to dance scholarship is that they can help us look at dance differently. Borrowing a term from Sinclair and Rockwell, we describe our visualizations and graphs as hermeneutica, interpretive toys that can change how a scholar looks at a particular cultural practice (dance in our case, text in theirs). Like them, we see our results "not as microscopes revealing the inner structure, but as augmentations adding to a history of interpretation" (Sinclair and Rockwell 2016: 101). Rich, qualitative descriptions of "body types" across multiple Javanese art-forms (puppetry, dance and dance-dramas) are available in previous dance scholarship and in ethnographies of Java. Dancers and scholars have developed evocative metaphors to describe the qualities of movement.

Our study constitutes the first systematic measurement of the differences among "body types" in Javanese dance but its goal is not to disprove previous scholarship. Our main interest is to explore how visualizations of our research data can be used to help scholars look at dance differently. We suggest a way forward for dance scholarship where interactive displays allow multiple ways of understanding dance, at the crossroads of ethnography, dance criticism and quantitative biomechanical analysis.

Bibliography

Bleeker, M. (2016), *Transmission in Motion: The Technologizing of Dance.* Abingdon: Routledge.

Rockwell, G., and Sinclair, S. (2016), *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. Cambridge: MIT Press.

Making Manuscripts Searchable: *DEx, a Database of Dramatic Extracts,* Digital Pub-lication, and Boutique Projects

Laura Estill
lestill@tamu.edu
Texas A&M University, United States of America

Too often, people lament that we have no evidence of what Shakespeare's readers thought. In reality, we have quite a bit of evidence, but it has been difficult to access because of the cost of travel and the barrier to access caused by early modern handwriting. This paper describes how *DEx: A Database of Dramatic Extracts* (beta version) contributes to the changing landscape of digital projects that help us better understand the early modern period. Rather than focusing on one person (too often, a canonical male literary figure), *DEx* instead makes a wide range of reader responses to early modern drama searchable, bringing together resources that are held in geographically distant archives, from Oxford's Bodleian Library to the Folger Shakespeare Library in Washington, D.C.

Dramatic extracts (those parts from plays that people copied into their notebooks) are important evidence because they tell what early modern audiences and readers took, literally and figuratively, from plays. Right now, this important evidence of reader response is difficult to access for multiple reasons: firstly, these extracts are often uncatalogued and therefore hard to locate. Secondly, they are dispersed across multiple archives and require travel funding to reach (see Ioppolo 2004). Finally, they are written in early handwriting, and so require paleographical training to read. This paper examines the ways in which *DEx*, a comparatively small project, links to existing resources and draws on community to ultimately provide people access to evidence of what Shakespeare's readers thought. (Spoiler: it turns out they thought about other playwrights more than Shakespeare in his lifetime!) *DEx* complements the Folger *Union First Line Index* and the *Catalogue of English Literary Manuscripts* by making the full text of dramatic extracts searchable and by not being constrained to selections written in verse (like the former) or by canonical authors (like the latter). *DEx* includes materials that are not catalogued in *CELM* or repository catalogues, and invites scholars to contribute relevant citations, transcriptions, or leads they have.

DEx's transcriptions are part of what makes it so valuable. Despite advances in OCR (Optical Character Recognition) technologies, there is no adequate program to automatically recognize handwritten text: particularly when it comes to historical documents written in scripts that we no longer use today, such as chancery hand or secretary hand. While digital paleography is an ongoing area of research, it tends to focus now on transcribing and describing texts by hand or on teaching paleography with digital tools (Stokes, 2014, 2015; Rehbein et al 2009, esp. 110-338; Fischer et al 2010; Hassner et al, 2013). Many digital projects that focus on early modern English texts are hand-transcribed and encoded: consider, for instance, the Folger Shakespeare Library's *Early Modern Manuscripts Online* and *Shakespeare's World*, *The Recipes Project* or *Bess of Hardwick's Letters.* Transcriptions in *DEx* are undertaken by a small community of scholars: this paper explains how our community currently works and the future collaborations we hope to undertake, as well as the possible avenues for extending the project after its forthcoming full launch.

In 2010, Paul Conway argued that "We are at the end of 'boutique' digital scanning projects for which the principal goal is ... extraordinary attention to the unique properties of each artifact" (76). This paper contends that with early modern manuscripts, "boutique" projects are one of the best ways forward. Compared to massive manuscript digitization projects like *British Literary Manuscripts Online*, *DEx* is a "boutique" project actually make texts searchable with transcription, which is always the result of paying attention to each manuscript as a "unique artifact." This paper discusses the challenges that come with curating a boutique project and the ultimate benefits of having a small site that emerges from a specific set of research questions.

Although small digitization or transcription projects can open up a vast field of research, they need to be findable and peer-reviewed in order to do so. I examine the obstacles to having *DEx* published by a traditional publisher, while questioning how to define publication for digital projects and the costs associated with creating and maintaining an open access site. Furthermore, I discuss how digital publishing must address the peer review needs of emerging scholars and provide an imprimatur and guarantee of quality for users. The final section of this paper discuss the role of Iter: Gateway to the Middle Ages and Renaissance and ReKN: Renaissance Knowledge Network in publication and peer review for *DEx: A Database of Dramatic Extracts*. This is an appropriate short paper for DH 2017 because it discusses a project that is in beta and active development, it engages the larger questions of how and why boutique digital projects can flourish and provide value to humanities scholarship, and it engages the theme of "Access/Accès" by focusing on collaboration, public-facing scholarship, and digital humanities publication. The paper focuses on a single case study: *DEx: A Database of Dramatic Extracts* and its community, which addresses a much-needed gap in scholarship by transcribing manuscripts that tell us what Shakespeare's audience and readers actually read.

Bibliography

Conway, P. (2010) "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly*. 61-79.

Fischer, F., Fritze, C., and Vogeler, G. (2010) *Codicology and Paleography in the Digital Age 2* (Norderstedt: Books on Demand).

Hassner, T, Rehbein, M., Stokes, P.A., and Wolf, L. (2013) "Computation and Palaeography: Potentials and Limits." *Dagstuhl Manifestos* 2: 14–35. doi:10.4230/DagMan.2.1.14

Ioppolo, G. (2004) "Switching on the World of Dramatic Manuscripts." *Shakespeare Studies* 32 (2004): 66-72.

Rehbein, M., Sahle, P., and Schassan, T. (2009) *Codicology and Paleography in the Digital Age* (Norderstedt: Books on Demand).

Stokes, P. (2014) "Describing Handwriting – again." In *Digital Palaeography: New Machines and Old Texts. Dagstuhl Reports* 4: 127–128. doi:10.4230/DagRep.4.7.112

Stokes, P. (2015) "Digital Approaches to Paleography and Book History: Some Challenges, Present and Future." *Frontiers in Digital Humanities.* 29 October 2015 http://dx.doi.org/10.3389/fdigh.2015.00005

Une approche de conception collaborative et d'exploitation des modèles ontologiques des données, facilement extensibles et compatibles avec le Web des Données Ouvertes (LOD) pour les Humanités Numériques (DH)

Hammou Fadili
fadili@msh-paris.fr
Fondation Maison des Sciences de l'Homme, France

Ahcène Ouguenoun
aouguenoune@adbi.fr
Accelerator Data & Business Intelligence, France

Résumé

Le but du présent article est de présenter une approche dont l'objectif est de mettre en place une plateforme générique, permettant la conception collaborative et l'exploitation de modèles ontologiques des données particuliers. Ils ont la particularité d'être facilement extensibles et compatibles avec le Web des Données Ouvertes *(Linked Open Data ou LOD),* destinée à être utilisée dans le domaine des humanités numériques *(Digital Humanities ou DH).* La démarche a été appliquée dans un premier temps à un instrument particulier : conception ontologique d'un wiktionnaire sémantique multilingue, multiculturel et multidisciplinaire des sciences humaines et sociales (SHS) afin d'une part de vérifier sur un exemple concret les fonctionnalités de la plateforme, et d'autre part de l'améliorer afin d'en faciliter la déclinaison à d'autres outils particuliers. En somme, le projet veut concevoir une fabrique de données intelligentes pour les humanités numériques *(Smart data factory for digital humanities)*; où la création des données suit un processus « cyclique », en deux étapes qui consistent (a) à créer directement dans la plateforme, par les experts du domaine, des données respectant toutes les normes exigées ; (b) à exploiter les données créées dans (a), en tant que données « expertes » validées, pour produire intelligemment et automatiquement, à partir de l'open data, de nouvelles données compatibles.

Introduction & motivation

L'objectif de ce travail vise la mise en place d'une plateforme centralisée d'aide à la conception collaborative de modèles ontologiques extensibles des données, facilitant la création et l'intégration des données interprétées et non ambiguës, dites données intelligentes *(Smart data)* au service des humanités numériques. Les contenus doivent être crées et générés sous forme de données structurées, sémantiquement annotées et liées, suivant des schémas de description bien adaptés. Dans notre cas, cela consiste à mettre en place un méta-modèle permettant de générer des modèles d'ontologies, des ontologies multilingues, multiculturelles et multidisciplinaires du domaine des SHS et une base de connaissances partagée et reconnue par des communautés de chercheurs.

Notre travail a été motivé par la fait que :

- Il n'existe pas suffisamment de données intelligentes, automatiquement exploitables, en SHS, à l'échelle internationale reflétant l'état de la coopération scientifique et culturelle entre la France et d'autres pays dont les concepts et lexiques pourraient évoluer de manières indépendantes
- Il y a un grand déséquilibre, d'un point de vue de la disponibilité des ressources numériques, entre le Français et les langues d'autres pays notamment celles des pays du sud
- Les dictionnaires multilingues, peu nombreux, sont souvent les résultats d'élaborations unilatérales
- Les traducteurs existants, également peu nombreux, ne prennent pas en compte tous les aspects liés aux contextes des définitions
- Les corpus multilingues pouvant constituer des sources de données sont également rares

- Les travaux sur les nouvelles technologies et la normalisation des données des langues de certains pays sont encore à leurs débuts

Dans cet article, la présentation de notre approche, se fera à partir de la description d'une contribution à la construction ontologique d'un Wiktionnaire sémantique multilingue, multiculturel et multidisciplinaire des SHS. Cette dernière est basée, entre autres, sur une adaptation et sur une extension de la plateforme collaborative « Mediawiki sémantique » existante afin qu'elle puisse prendre en compte nos modèles ontologiques des données. Sa construction devrait permettre aux chercheurs d'échanger et de partager des connaissances dans le domaine des SHS et cela quelques soient leurs spécialités, leurs langues et leurs lieux géographiques de travail et/ou de résidence. Sa réalisation a intégré en plus, des normes et des protocoles bien spécifiques, en vue de son intégration dans le Web de Données Ouvertes (LOD).

Notre approche

Modèles des données

La conception de la structure de l'ontologie et donc du Wiktionnaire repose sur des correspondances entre les éléments de départ dans leurs contextes pour une langue source et les éléments d'arrivée dans leurs contextes pour une/des langue(s) cible(s) selon un sous-ensemble du schéma de la norme ISO1951. Pour simplifier, on va considérer les langues par paires. Donc, pour définir le modèle, on doit prendre en compte le fait qu'une entrée A_k dans une langue source peut avoir plusieurs sens et donc plusieurs traductions B_1, ..., B_j, ...B_m dans une langue cible. Cette même entrée A_k peut être définie avec plusieurs éléments A_1, ..., A_i, ...A_n du schéma du dictionnaire (cf. FIG. 2) qui peuvent êtres à leurs tour des entrées dans la même langue source et par conséquent, peuvent avoir plusieurs sens dans cette même langue source et plusieurs traductions dans la langue cible (FIG. 1). Notons que selon le sens de la traduction une langue source peut devenir cible et réciproquement.

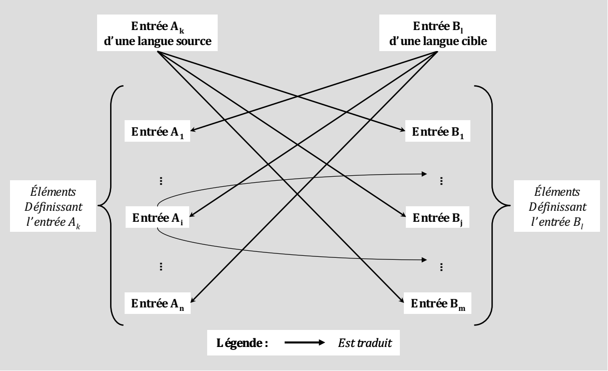

Entrée A_k d'une langue source
Entrée B_l d'une langue cible
Entrée A_1
Entrée A_i
Entrée A_n
Entrée B_1
Entrée B_j
Entrée B_m
Éléments Définissant l'entrée A_k
Éléments Définissant l'entrée B_l
Légende : Est traduit

Figure 1. Extrait du schéma du dictionnaire des SHS

Ce schéma, récapitulant ce qui a été décrit précédemment, montre la complexité des renvois entre une langue source et une langue cible, ayant des spécificités différentes. Ce qui nécessite la définition d'un schéma spécifique dans chaque langue ainsi que la mise en place d'un système de gestion des correspondances d'une manière automatique. Nous pouvons procéder tout d'abord par une première simplification du problème de départ, qui consiste à associer à une entrée source (mot, locution, etc.) un ou plusieurs sens (définitions) qui renvoient à une ou plusieurs entrées cibles ; puis revenir du terme traduit, pris cette fois-ci comme entrée source. Ce procédé a été pris en charge par la mise en place d'un système de guidage d'aide à la génération et à la définition des correspondances entre les entrées, leurs définitions et leurs traductions dans les différentes langues. L'utilisateur peut modifier ou valider les suggestions du système pour compléter les fiches des entrées suivant les critères suivants (figure 2) :

- La définition d'une entrée se fait par la description d'une fiche suivant un format structuré déterminé par le schéma : contextes, définitions, relations sémantiques, traductions, indications grammaticales, parlers, etc.
- la signification attribuée à une entrée dépend d'un contexte de définition. Ce dernier est décrit par un ensemble fini et connu de paramètres contextuels des aspects: temporels, géographiques, disciplinaires, culturels, linguistiques, etc.
- Les relations entre les termes, se fait par le biais de relations sémantiques telles que : la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, l'isonymie, la conversion, etc.

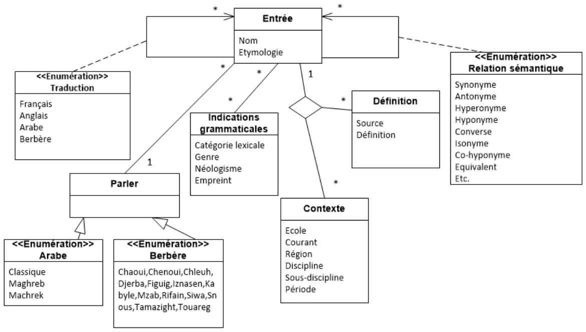
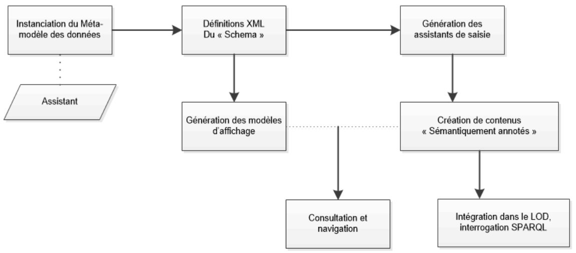


Figure 2.- Modèle du schéma du Wiktionnaire

Plateforme

L'architecture mise en place pour la réalisation repose sur une conception générique, capable de prendre en charge plusieurs modèles de données « *schémas de données*» et plusieurs langues ; et reste flexible et facilement extensible à d'autres schémas et généralisable à d'autres langues. L'implémentation a été réalisée, suivant 4 modules fondamentaux, assurant la gestion de la totalité du workflow (figure 3).


Instanciation du Méta-modèle des données
Assistant
Définitions XML Du « Schema »
Génération des modèles d'affichage
Génération des assistants de saisie
Création de contenus « Sémantiquement annotés »
Consultation et navigation
Intégration dans le LOD, interrogation SPARQL

Modules du processus de l'application Wiktionnaire sémantique des SHS

Figure 3. Processus général

Le processus consiste en l'instanciation du méta-modèle des données et la génération des différents schémas d'utilisation. Cela se fait par des définitions XML (qu'on peut générer grâce à des assistants dédiés) permettant la génération des modèles des données, des assistants de saisies et d'annotations ainsi que des modèles d'affichage. Le processus intègre également des modules d'exploitation et d'intégration des données dans le LOD (import/export des triplets RDF), ainsi qu'module de consultation et de navigation (exposition d'un point de terminaison SPARQL).

Exploitation

Après les phases de tests et de validation des modèles implémentés, l'application est actuellement en production. Elle est alimentée par un réseau d'enseignants chercheurs répartis suivant les disciplines et les langues de leurs spécialités. Elle mobilise de plus en plus de chercheurs et comporte actuellement plusieurs entrées dans plusieurs langues et disciplines. Il est également important de noter que certains éléments du wiktionnaire ont été enrichis, grâce à des requêtes SPARQL bien paramétrées sur du LOD (comme DBpedia).

Conclusion

La plateforme permet, d'une part, de créer des contenus « Intelligents » directement dans la plateforme, d'autre part, d'utiliser les données créées pour les confronter avec des sources de données externes de l'Open Data et du Web de données (Linked open Data), pour générer de nouvelles données Intelligentes. La solution développée étant générique, son extension et son adaptation à d'autres domaines et à d'autres langues est une tâche facile. Une des perspectives de ce travail consiste à encourager la dynamique actuelle, afin que la plateforme puisse devenir une référence dans le domaine des ressources ontologiques et dictionnairiques au service des Humanités Numériques.

Bibliographie

Berner-Lee, T. (2010). "Open, Linked Data for a Global Community." presented at the Gov 2.0 Expo 2010, May 26.

Harris, K. (2012). "Explaining Digital Humanities in Promotion Documents." Journal of Digital Humanities 1, no. 4

Grupo Tragsa. (2013) Smart Open Data. http://www.smartopendata.eu/

Linked Data Group (2009). Linked Data: Connect Distributed Data Across the Web. http://linkeddata.org/

Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts

Frank Fischer
frafis@gmail.com
National Research University
Higher School of Economics, Russia

Mathias Göbel
goebel@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Dario Kampkaspar
kampkaspar@hab.de
Herzog August Library, Germany

Christopher Kittel
contact@christopherkittel.eu
University of Graz, Austria

Peer Trilcke
trilcke@uni-potsdam.de
University of Potsdam, Germany

Introduction and related works

In recent years, the application of network analysis methods to literary texts has evolved into an independent research field of digital literary studies. Methods for the automated extraction of network data (named entity recognition, co-reference resolution) and their evaluation are of particular importance (Elson et al. 2010; Park et al. 2013; Agrarwal et al. 2013; Fischer et al. 2015; Waumans et al. 2015; Jannidis et al. 2016). Based on the data obtained, several types of analyses were developed: an empirical, quantitative and hierarchical description of literary characters (Jannidis et al. 2016), corpus-based analyses exploring options for historical periodisation of literature (Trilcke et al. 2015) and types of aesthetic modelling of social formations in and by literary texts (Stiller et al. 2003; Stiller & Hudson 2005; Trilcke et al. 2016).

What has been neglected so far (although already suggested by Moretti 2011) is the application of network analysis as a tool for quantitative plot analysis. In fact, current approaches in the field of literary network analysis are not suited to gaining insights into the plot development of literary texts (Prado et al. 2016). The sequential dimension of literary texts, as a consequence of their temporality, usually remains in the dark: what is extracted, visualised and analysed are *static* networks. Plot, however, is essentially a concept supposed to theoretically encompass the temporality of narrative (as well as dramatic) texts: "the repeated attempts to redefine parameters of plot reflect both the centrality and the complexity of the temporal dimension of narrative" (Dannenberg 2005: pp. 435). Plot can be understood as a concept for the description of the "progressive structuration" (Kukkonen 2013: §4) of literary texts.

Research objective: plot analysis

Attempts to further develop literary network analysis towards a quantitative plot analysis must consider the temporal dimension in the modelling of their research objects. The structure of a literary text is to be modelled as a changing sequence of network states. It is only through looking into these network dynamics that we can discuss network-analytic approaches for a quantitative plot analysis.

Following Prado et al. 2016, we are currently extending our research on literary networks (Trilcke 2013; Fischer et al. 2015ff.; Fischer et al. 2015; Trilcke et al. 2016) to the analysis of *progressive* structuration. Our goal is to examine whether (and with what kind of limitations) we can flesh out an operationalisation for the plot analysis of literary texts. By doing so, we are, of course, not trying to replace the semantically rich and versatile concept of 'plot' with a quantitative and thus reductionist concept. Rather, we will show that certain aspects of what is commonly discussed within the framework of plot analysis can be retraced by means of a computer-based analysis of network dynamics.

The visualisation of dynamic graphs, as is common in other domains (Pohl et al. 2008; Frederico et al. 2011), has recently been transferred to literary networks (Xanthos et al. 2016). While it may come in useful when close-reading a text and for didactic purposes, it is unsatisfactory when it comes to an actual corpus-based analysis. There are no canonical methods to help us compare network visualisations generated automatically by force-directed graph drawing algorithms. The reception of dynamic visualisations just does not offer practical analytical means. From a dedicated distant-reading kind of perspective, the calculation of dynamic network metrics and their statistical processing is much more promising as it offers options to describe general characteristics of networks from a larger corpus as well as to compare specific formal types of networks within the corpus.

Measuring dynamic literary networks

A number of basic global measures for the analysis of dynamic networks (i.e., size, density, homogeneity in the distribution of ties, rate of changes in nodes, rate of changes in ties) has been discussed by Carley (2003: pp. 135–36). In addition, Prado et al. 2016 recently suggested the application of actor-oriented measures, especially centrality indices, for the reconstruction of plot development. We are currently applying these measures to our own corpus, consisting of 465 German language plays. We also developed a set of other measures with recourse to traditional theoretical concepts for the description of specific phenomena of plot development, especially with regard to types of dramatic expositions (Pfister 1977: pp. 124–36), the "classical" act structure of tragedy and the composition principle of main and secondary plots (Pfister 1977: pp. 286–89). Calculation of these measures was implemented in our own network analysis tool *dramavis* (Kittel/Fischer 2017).

Event-based measures

In general, two types of measures can be distinguished for describing dramatic texts as dynamic networks: event-based and progression-based measures. Event-based measures are used to identify or characterise a particular point in time within a drama. In this context we developed an *all-in index*, a value that identifies the point in time at which all characters have occurred at least once in a drama (see Figure 1).

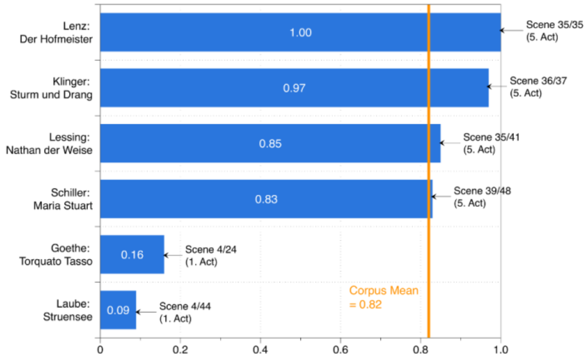


Figure 1: All-in index for 6 selected plays

The *final-scene-size* value characterises a specific point in time, in this case the last scene of a drama. It determines the percentage of all characters of a drama which appear on stage in the last scene. This value shows characteristic differences between dramatic subgenres, especially when comparing tragedies and comedies (see Figure 2).

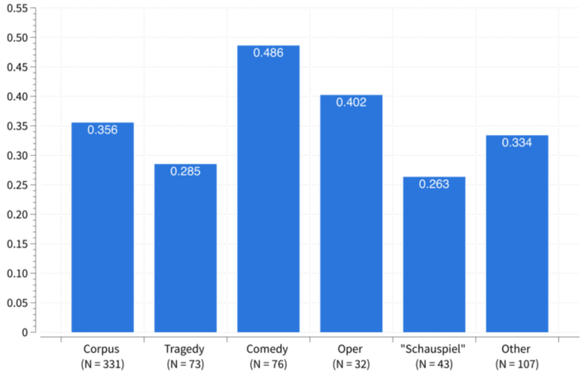


Figure 2: Final-scene-size index (mean values of entire corpus and subgenres)

Progression-based measures

While event-based measures allow assumptions about a particular state/point in time of the dynamic network, progression-based measures allow a general characterisation of the transformation of a dramatic network. In this regard, we introduced a measure we call the *drama-change rate*. The basis of our calculation is a modified Levenshtein distance, which only takes into account insertions ("add character") and deletions ("delete character"). In each case, we compare characters present in two consecutive scenes, eventually resulting in what we call the *segment-change rate* (Figure 3).

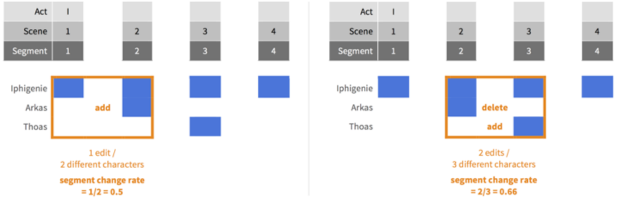


Figure 3: Calculation of the segment-change rate (example)

The sum of all *segment-change rates* of a drama divided by the number of all *segment-change rates* is what we call the *drama-change rate*. The development of this rate can be represented in a chart, which due to its shape we tentatively called the *beat chart* (Figure 4).

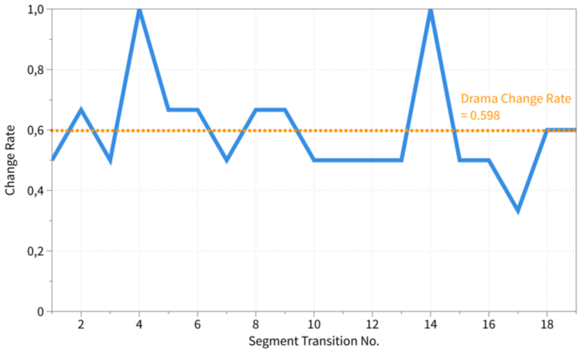


Figure 4: Beat chart for Goethe's play "Iphigenie auf Tauris" (1787)

Having calculated the *drama-change-rate* value for the entire corpus, we can start to compare a larger set of dramas with each other (Figure 5). It becomes evident that our corpus does not exhibit a clear trend along the timeline. Instead, we witness the emergence of characteristic types of dramas, which differ characteristically from the other texts in our corpus.

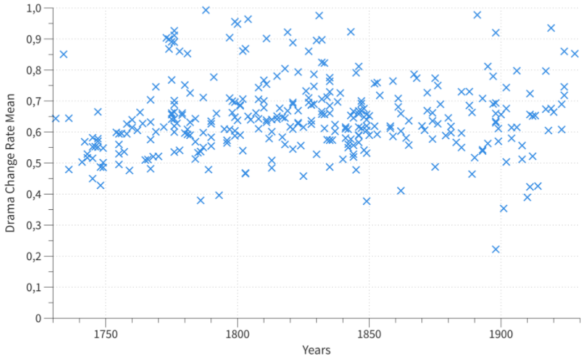


Figure 5: Historical distribution of drama-change rates in the DLINA corpus

On the one hand, we can identify dramas exhibiting a high *drama-change rate*, i.e., highly dynamic dramas with a constant alternation of characters on stage (Figure 6). On the other hand, low-dynamic dramas can be identified, with only small changes taking place between scenes (Figure 7).

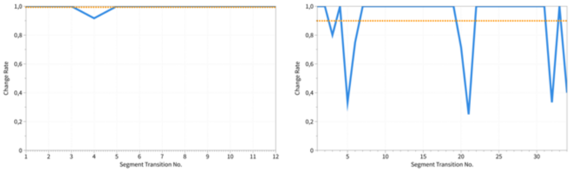


Figure 6: Two examples of highly dynamic dramas – left: Goethe's "Egmont" (1788); right: Lenz's "Der Hofmeister" (1774)

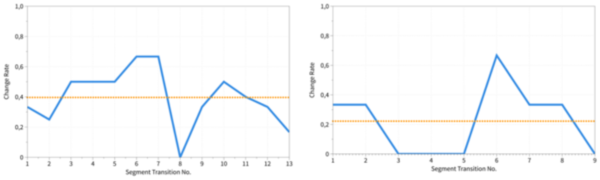


Figure 7: Two Examples for low-dynamic dramas – left: Goethe's "Der Bürgergeneral" (1793), right: Rilke's "Ohne Gegenwart" (1898)

A further option for the comparative analysis of our oscillatory *beat charts* is the analysis of the standard deviation of all *segment-change rates* of a drama. We can again distinguish two particularly striking types: On the one hand, there are dramas with a high standard deviation, indicating that extensive changes of characters alternate with small changes; we can call them high-dynamic dramas (Figure 8). On the other hand, there are dramas showing a low standard deviation, so the change on stage takes place in the same rhythm, we call these texts low-dynamic dramas (Figure 9).

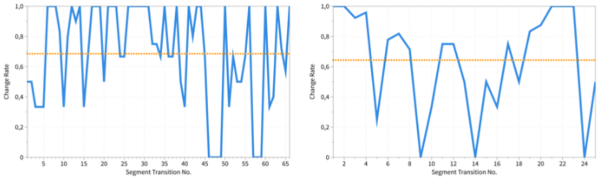


Figure 8: Two Examples for high-dynamic dramas – left: Benkowitz's "Die Jubelfeier der Hölle" (1801), right: Goethe's "Faust I" (1808)

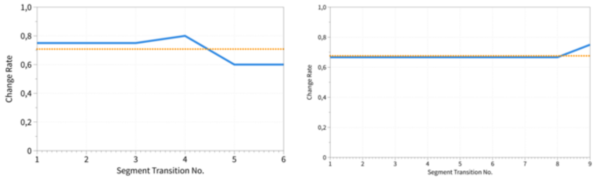


Figure 9: Two Examples for low-dynamic dramas – left: Schnitzler's "Anatol" (1893), right: Schnitzler's "Der Reigen" (1902)

The preceding cases each describe characteristic deviations regarding the *drama-change rate* of groups of texts. In addition, a 'normal type' of drama can be identified, which corresponds to the mean value for the corpus both in terms of arithmetic mean and standard deviation (Figure 10).

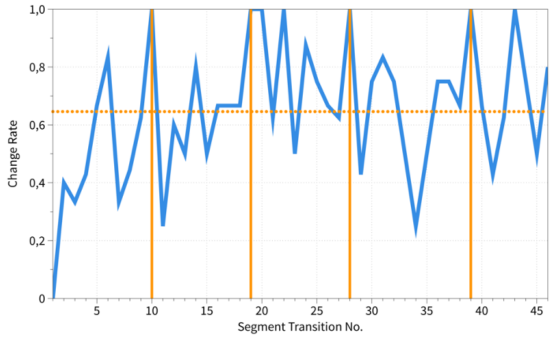


Figure 10: Beat chart for Ganghofer's "Der Herrgottschnitzer von Ammergau" (1880)

It also appears that strong upward shifts, accompanied by a complete exchange of characters on stage, often coincide with act boundaries (vertical orange lines), which is in accordance with historical conventions of dramatic composition.

Summary and further research

Our research on dynamic networks provides basic components for a quantitative analysis of the progressive structuration of dramatic texts. Future research will have to develop and evaluate additional measures and it will be decisive to hold interpretations of these measures against the backdrop of historical drama poetics.

Bibliography

Agarwal, A., Corvalan, A., Jensen, J., Rambow, O. (2012). Social Network Analysis of Alice in Wonderland. *Proceedings of the Workshop on Computational Linguistics for Literature.* Montréal, pp. 88–96, http://www.aclweb.org/anthology/W12-2513 [accessed 27 Mar 2017].

Carley, K. M. (2003). Dynamic Network Analysis. Breiger, R., Carley, K. M., Pattison, P. (eds.): *Dynamic Social Network Modeling and Analysis.* Workshop Summary and Papers. Washington D. C., pp. 133–45, http://www.nap.edu/read/10735/chapter/9 [accessed 27 Mar 2017].

Dannenberg, H. (2005). Plot. Herman, D., Jahn, M., Ryan, M.-L. (eds.): *The Routledge Encyclopedia of Narrative Theory.* London: Routledge, pp. 435–39.

Elson, D. K., Dames, N., McKeown, K. R. (2010). Extracting Social Networks from Literary Fiction. *Proceedings of ACL 2010.* Uppsala, pp. 138–47, http://dl.acm.org/citation.cfm?id=1858696 [accessed 27 Mar 2017].

Federico, P., Aigner, W., Miksch, S., Windhager, F., Zenk, L. (2011). A Visual Analytics Approach to Dynamic Social Networks. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW).* Graz, http://publik.tuwien.ac.at/files/PubDat_198995.pdf [accessed 27 Mar 2017].

Fischer, F., Göbel, M., Kampkaspar, D., Trilcke, P. (2015). Digital Network Analysis of Dramatic Texts. *Digital Humanities 2015. Conference Abstracts.* University of Western Sydney, http://dh2015.org/abstracts/xml/FISCHER_Frank_Digital_Network_Analysis_of_Dramati/FISCHER_Frank_Digital_Network_Analysis_of_Dramatic_Text.html [accessed 27 Mar 2017].

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., Trilcke, P. (2015ff.). [Blog] dlina – Digitally-Driven Literary Network Analysis (of Dramatic Texts). https://dlina.github.io/ [accessed 27 Mar 2017].

Jannidis, F., Reger, I., Krug, M., Weimer, L., Macharowsky, L., Puppe, F. (2016). Comparison of Methods for the Identification of Main Characters in German Novels. *Digital Humanities 2016. Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 578–82 http://dh2016.adho.org/abstracts/297 [accessed 27 Mar 2017].

Kittel, C., Fischer, F. (2017). dramavis (v0.3). GitHub repo: https://github.com/lehkost/dramavis [accessed 27 Mar 2017].

Kukkonen, K. (2013). Plot. Hühn, P. et al. (eds.): *The Living Handbook of Narratology.* Hamburg. http://www.lhn.uni-hamburg.de/article/plot [accessed 27 Mar 2017].

Moretti, F. (2011). Network Theory, Plot Analysis. *Stanford Literary Lab Pamphlets.* No. 2, 1 May 2011, http://litlab.stanford.edu/LiteraryLabPamphlet2.pdf [accessed 27 Mar 2017].

Park, G.-M., Kim, S.-H., Cho, H.-G. (2013). Structural Analysis on Social Network Constructed from Characters in Literature Texts. *Journal of Computers* 8.9, pp. 2442–47, http://ojs.academypublisher.com/index.php/jcp/article/view/jcp080924422447/7672 [accessed 27 Mar 2017].

Pfister, M. (1977). Das Drama. Theorie und Analyse. München: Fink.

Pohl, M., Reitz, F., Birke, P. (2008). As Time Goes by. Integrated Visualization and Analysis of Dynamic Networks. *AVI 2008 – Proceedings of the Working Conference on Advanced Visual Interfaces.* Neapel, pp. 372–75, http://doi.acm.org/10.1145/1385569.1385636 [accessed 27 Mar 2017].

Prado, S. D., Dahmen, S. R., Bazzan, A. L. C., Mac Carron, P., Kenna, R. (2016). Temporal Network Analysis of Literary Texts. *Advances in Complex Systems (ACS)* **19**(3), pp. 1–19, https://arxiv.org/pdf/1602.07275 [accessed 27 Mar 2017].

Rochat, Yannick (2014). Character Networks and Centrality. Thèse de Doctorat, Lausanne, https://infoscience.epfl.ch/record/203889/files/yrochat_thesis_infoscience.pdf [accessed 27 Mar 2017].

Stiller, J., Nettle, D., Dunbar, R. I. M. (2003). The Small World of Shakespeare's Plays. *Human Nature* 14, pp. 397–408, https://www.staff.ncl.ac.uk/daniel.nettle/shakespeare.pdf [accessed 27 Mar 2017].

Stiller, J., Hudson, M. (2005). Weak Links and Scene Cliques Within the Small World of Shakespeare. *Journal of Cultural and Evolutionary Psychology* 3, pp. 57–73.

Trilcke, P., Fischer, F., Göbel, M., Kampkaspar, D. (2015). 200 Years of Literary Network Data [blogpost], https://dlina.github.io/200-Years-of-Literary-Network-Data/ [accessed 27 Mar 2017].

Trilcke, P., Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C. (2016). Theatre Plays as 'Small Worlds'? Network Data on the History and Typology of German Drama, 1730–1930. *Digital Humanities 2016. Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 385–87 http://dh2016.adho.org/abstracts/407 [accessed 27 Mar 2017].

Trilcke, P. (2013). Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft. Ajouri, P. et al. (eds.): *Empirie in der Literaturwissenschaft.* Münster: mentis, 201–247.

Waumans, M. C., Nicodème, T., Bersini, H. (2015). Topology Analysis of Social Networks Extracted from Literature. *Plos One*, 3 June 2015, http://dx.doi.org/10.1371/journal.pone.0126470 [accessed 27 Mar 2017].

Xanthos, A., Pante, I., Rochat, Y., Grandjean, M. (2016). Visualising the Dynamics of Character Networks. *Digital Humanities 2016. Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 417–19 http://dh2016.adho.org/abstracts/407 [accessed 27 Mar 2017].

Transcription of Encoded Manuscripts with Image Processing Techniques

Alicia Fornés
afornes@cvc.uab.es
Universitat Autònoma de Barcelona, Spain

Beáta Megyesi
beata.megyesi@lingfil.uu.se
Uppsala University, Sweden

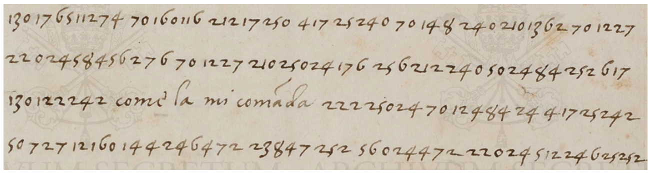
Joan Mas
jmas@cvc.uab.es
Universitat Autònoma de Barcelona, Spain

Introduction

Historical hand-written manuscripts are an important source of information for our cultural heritage, and automatic processing of these can help exploring their content faster and easier.

A special type of hand-written manuscripts that are relatively common in archives and libraries are encrypted, secret documents, so called ciphers. Ciphers may contain and hide important information for the history of science, religion or diplomacy and therefore shall be decrypted and made accessible. The automatic decryption of historical hand-written ciphers is the main focus of the project *DECODE: Automatic decoding of historical manuscripts.* In order to reveal the content of these secret messages, we collect and digitize ciphertexts and keys from Early Modern times, build a database, and develop software tools for (semi-)automatic decryption by cross-disciplinary research involving computer science, language technology, linguistics and philology.

Ciphers use a secret method of writing, often by transposition or substitution of characters, special symbols, already existing alphabets, digits, or a mixture of these. The encoded sequences are usually meticulously written and often segmented character by character to avoid any kind of ambiguity for the receiver to be able to decode the content, but continuous writing of some sequences where the symbols are connected also exists. In addition, the cipher sequences might be embedded in cleartext, i.e. texts in a known natural language, as illustrated in the picture below (Archivio Segreto Vaticano, 2016).



13017 6511274 7016 0116 2121 7250 417 25240 70 148 240 2101362 701227
2202458456276 70 1227 2102502417 6 25 62122 240 502484 252 617
13012 2242 comé la mi comada 22225024 70124 84 24 417 25242
50 727 12160 1442 46472 23847 252 5602447 2 22024 51224 625252

The first step for deciphering and making accessible the secret writing is their digitization and transcription. Transcription can be performed either by hand where a person types in the encrypted text symbol by symbol, or by (semi-)automatic means with a possible post editing by manual validation. Manual transcription is time-consuming and expensive, and prone to errors. Automatic methods applied to ease the transcription process are preferable. However, image processing techniques developed so far for historical text manuscripts, such as the ones from the project *TransScriptorium*, are not fully adequate for dealing with encrypted documents for several reasons. First, the transcribing system cannot benefit from any lexicon or language model because the key is, a priori, unknown. Consequently, the use of an optical model alone is prone to errors, especially when there are ambiguities in the shape of digits/characters. Second, many ciphers contain a mixture of plaintext and encrypted text (ciphertext), which requires specifically adapted handwriting recognition methods. Third, the arcane nature of the symbols used calling for semiotic analysis, which requires the study of techniques closer to hand-drawn symbol recognition rather than handwriting recognition ones.

In this paper, we study the feasibility of the current image processing techniques in order to digitize ciphers by recognizing and transferring the symbols into a computer-readable format. For this purpose, we present a semi-automatic transcription method based on Deep Neural Networks, followed by a manual validation. We compare the results with a complete manual transcription, and analyze the human time effort of the two scenarios.

Image Processing Methodology

The handwriting recognition system has the following steps: First, each document has been binarized, deskewed, and the text lines have been segmented using projection profiles. The images of the text lines are the input of the Multi-Dimensional Long Short-Term Memory Blocks Neural Networks (MD-LSTMs) (Graves and Schmidhuber, 2008; Voigtlaender and Doetsch, 2016). Contrary to previous techniques applied for recognition (e.g. HMMs), MD-LSTMs obtain good results without the need of computing feature vectors from the image.

For each text line, the output of the network is a sequence of digits. The system also detects when a digit has a dot above or below. Whenever the confidence of the system when transcribing a certain digit is low, the symbol "?" is used. This denotes that an expert user must check it.

In this work, the networks were trained using 15 cipher pages from six different ciphers (with six different handwritings) in order to learn the handwriting style variability. For validation set, we used 5 cipher pages from the same ciphers as in the training set but different pages. It must be noted that the amount of training pages is not enough for the transcription of cleartext, so all text words appearing in the document are denoted as "x". The transcription must be performed by an expert.

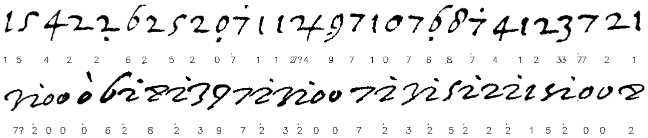
Finally, and with the aim of improving the visualization of the results and facilitating the posterior validation and correction task, we used force alignment between the result of the neural network and the input image.

Manual vs. Automatic Transcription and Correction

For the tests, we chose 14 new unseen cipher pages, of which 12 pages were taken from four ciphers in the training set, and two pages came from two new, previously unseen ciphers, which means that these handwriting styles have not been learned during training. In this way, we can also analyze the generalization and scalability degree of the method for new handwriting styles.

To compare the speed of automatic versus manual transcription in a fair way, each manuscript was manually transcribed by one person, whereas the output from the automatic transcription was corrected and validated by a different person.

In the manual transcription, the transcriber opened the image of the cipher, and transcribed it symbol by symbol in a text file. Contrary, for validation, the transcriber opened the output from the automatic transcription as a picture where the cipher page was segmented line by line and the suggested transcription was reproduced below. As it can be observed in the figures below, the symbol "?" appears when the system is not confident on the transcribed digit. Also, if the system detects a dot above the digit, then the transcription also contains the dot.



The results obtained by manual transcription and validated transcription from automatic output were compared and shown in Table 1. In average, the automatic system transcribes the digits with an average accuracy of 88 %. However, one of the ciphers (Francia_18_3_233r), written by a writer whose handwriting was not represented in the training set, was more difficult to the system to automatically transcribe and accuracy decreased to 61%.

Cipher	No. of lines per page	Manual (mins)	Validation (mins)	Accuracy automatic	Manual mins/line	Validation mins/line
Francia_4_1_221r	3	5	4	92%	1.67	1.33
Francia_6_1_236r	31	50	47	92%	1.61	1.52
Francia_18_2_206v	24	45	41	81%	1.88	1.71
Francia_18_3_233r	20	45	30	61%	2,25	1.50
Francia_64_2_040v	24	25	30	92%	1.04	1.25
Francia_64_4_056v	26	20	52	87%	0.77	2.00
Francia_64_5_060v	25	20	26	94%	0.80	1.04
Francia_64_6_064v	16	10	13	94%	0.63	0.81
Spagna_423_2_297r	8	15	4	98%	1.88	0.50
Spagna_423_3_300v	2	3	3	74%	1.50	1.50
Spagna_423_4_374r	10	15	10	85%	1.50	1.00
Spagna_423_6_388v	21	35	15	95%	1.67	0.71
Spagna_423_7_391r	13	15	8	97%	1.15	0.62
Spagna_423_9_491v	21	25	20	93%	1.19	0.95
Average	**17.43**	**23.43**	**21.6**	**88%**	**1.39**	**1.17**

Table 1. Summary of results per line and cipher given the time in minutes for manual transcription, as well as the validation and correction of automatic transcription; the accuracy of automatic transcription, and the average rate of manual transcription and validation per line. Rows in red color denote those where the manual transcription is faster.

The results show that in most cases manual transcription is 15% slower on average compared to the automatic transcription with post-editing if the accuracy of the image processing is above 90%. When accuracy is lower, validation time usually increases because the more transcription errors we find, the more effort it takes to localize both the wrong symbol(s) in the picture and in the transcription file. For each error, the user usually starts to read the line from the beginning.

It is also noteworthy that we do not count the time it takes to prepare and train the automatic transcription models, including the preprocessing of the images (cut the margins, clean the bleed through, etc.) and the time for training the models. We also noted that the validators would have benefited from a user-friendly transcription tool where transcription suggested by the model was aligned with the original symbol in the picture. However, the automatic transcription clearly helped to differentiate between two different symbols written similarly, thereby helping the user to identify the symbol set represented in the cipher.

Conclusion

We have shown that image processing can be used as base for transcription followed by a post-processing step with user validation and correction. Even though image processing techniques need to be trained today on individual handwritings to reach high(er) accuracy, they might be of great help to identify the symbol set represented in the manuscript and to make clear distinctions between symbols, hence can be used as a support tool for the transcriber.

In this work, we focused on ciphers without any esoteric or other symbol sets, which might be more difficult for an automatic recognition system. Also, we have identified only cipher sequences; cleartext was only detected without any further transcription.

In the future, we would like to test combining image processing and automatic decryption in one step to skip the time-consuming transcription step and create synergy effects as both image processing and automatic decryption tools rely on language models that could be used simultaneously. Another alternative is image processing for validation of the manual transcription, which might be an interesting alternative to investigate in the future.

Acknowledgement

This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the Swedish Research Council DECODE project grant E0067801, and the Ramon y Cajal Fellowship RYC-2014-16831.

Bibliography

Segr.di.Stato/Portogallo/1A/16v@2016. Archivio Segreto Vaticano. The picture has been reproduced by the kind permission of Archivio Segreto Vaticano, all rights reserved.

Frinken, V., and Bunke, H (2014). Continuous Handwritten Script Recognition. In *Handbook of Document Image Processing and Recognition.* Springer-Verlag, 2014.

Graves, A., and Schmidhuber, J. (2008). “Offline handwriting recognition with multidimensional recurrent neural networks”. *Neural Information Processing Systems,* 2008.

Voigtlaender, P. and Doetsch, H (2016). Ney. Handwriting Recognition with Large Multidimensional Long Short-Term Memory Recurrent Neural Networks. *International Conference on Frontiers in Handwriting Recognition*, 2016.

The Shape of History: Reimagining Nineteenth-Century Data Visualization

Caroline Foster
Georgia Institute of Technology
United States of America
cfoster2@gatech.edu

Adam Hayward
adam.hayward@gatech.edu
Georgia Institute of Technology
United States of America

Svyatoslav Kucheryavykh
Georgia Institute of Technology
United States of America

Angela Vujic
Georgia Institute of Technology
United States of America

Maninder Japra
Georgia Institute of Technology
United States of America

Shivani Negi
shivani.negi@gatech.edu
Georgia Institute of Technology
United States of America

Lauren Klein
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology
United States of America

Introduction

In the mid-1850s, American educator and editor Elizabeth Peabody (1804-1894) set off from Boston to ride the rails. She traveled as far north as Rochester, NY; as far west as Louisville, KY; and as far south as Richmond, VA, in order to promote the textbook she had recently published, *A Chronological History of the United States* (1856). Along with her suitcase, Peabody traveled with a large fabric roll, which, when unrolled, displayed a grid-like array of colored squares that represented the major events in U.S. history. In the nineteenth-century version of a product demo, Peabody would arrange the “painted centuries,” as she called them, on the floor, and invite potential textbook adopters to sit around the charts and contemplate the colors and patterns that they perceived (9).

Although not described in terms of visualization--the term did not enter common parlance until the early twentieth century--Peabody’s ideas about the uses of her charts anticipate many of the benefits associated with visualization today: the ability to “offload” mental processing “from cognitive to perceptual systems,” to “enhance” pattern recognition through “abstraction and aggregation,” and, crucially, to interact with and potentially “manipulate” the visualization itself (Card et al. 1999, 16). For Peabody did not only imagine that her readers would interpret the “data” presented on her charts; she also intended for them to create charts of their own. To this end, Peabody also sold workbooks of blank charts, so that students could read each chapter of her textbook, and then convert the list of events that followed into color and position, according to her visual scheme.

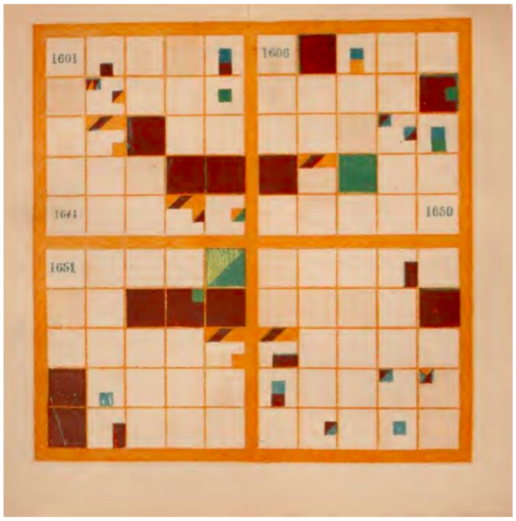


Figure 1: Peabody’s visualization of the significant events of the seventeenth century. In *A Chronological History of the United States, arranged with plates on Bem’s principle* (New York: Sheldon, Blakeman, 1856).

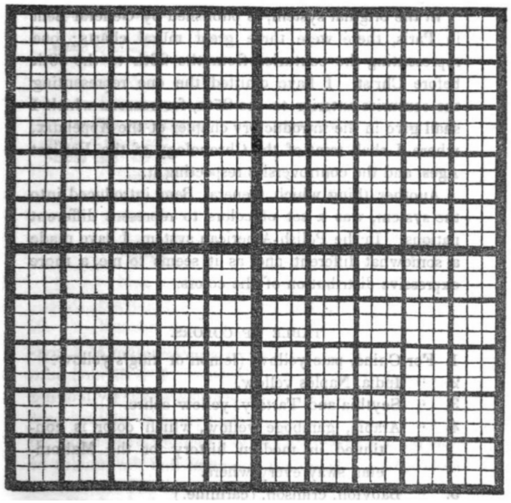


Figure 2: A blank chart included in The Polish-American System of Chronology: Reproduced, with some modifications, from General Bem’s Franco-Polish method (New York: G.P. Putnam, 1850).

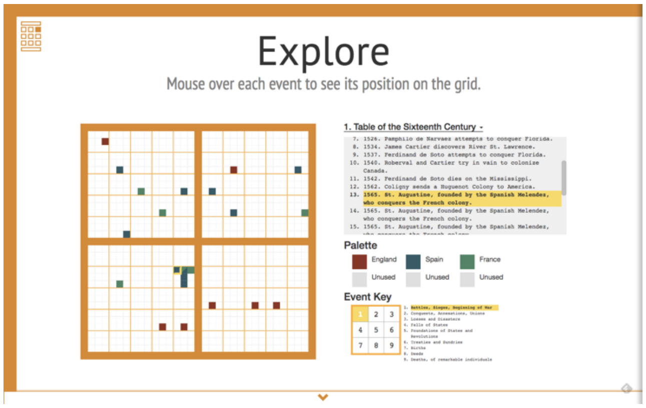
Project Overview

Drawing from recent digital humanities work relating to historical fabrication (e.g. Elliott et al. 2012, Sayers 2015), as well as from our own previous explorations of historical visualization techniques (e.g. Foster et al. 2016), we set out to recreate and enhance Peabody’s pioneering visual design. In particular, we focused on Peabody’s ideas about interaction and interpretation, since her ideas about the tripartite relation between data, text, and image-- and the role of the reader in translating between each-- speak directly to current debates in the digital humanities about the importance of acknowledging data as “capta” (Drucker 2011), and of recognizing the role of individual interpretation in both the design and reception of visualizations (Posner 2016). In our project, we focused first on reimagining Peabody’s original interaction for the web, employing current information visualization research to suggest techniques for emphasizing the interrelation between the data and their visual display. We then began a project to recreate the floor-sized version of Peabody’s chart using physical computing materials, so as to further explore the embodied aspects of Peabody’s visualization scheme. In the following sections, we describe the design choices involved in each recreation-- the digital and the physical-- with particular attention to how we sought to amplify Peabody’s ideas about interaction, interpretation, and embodiment through our reimagined interfaces.

The Shape of History: Reimagining Interaction and Interpretation for the Web

The Shape of History website represents the culmination of a year-long iterative design process. From Peabody’s original textbook, we distilled four conceptual modes of interaction: an “explore” mode, designed to explain to novice users how to interpret her charts, and how to translate between text and image; a “lesson” mode, designed to allow users to create their own charts, drawing upon Peabody’s original data; a “compare” mode, designed to call attention to how choices in visual display affect the charts’ ultimate interpretation; and a “play” mode, intended to facilitate the most open-ended form of interaction and expression. To implement the site, we employed a combination of HTML5, CSS, and JavaScript, including Bootstrap.js for site structure, jQuery for navigation and site-level interaction, and D3.js and two.js (along with custom JavaScript) for the visualization components.

At each juncture, we considered how to enhance Peabody’s original designs and interactions. For instance, when recreating the grid that would serve as the primary typographical form, we remained faithful to the original design and color palette, while adding additional minor grid lines (in light gray) so that users would know where to click (White 2011). In order to emphasize the relation between text and image, an important feature of both the “explore” and “compare” modes, we added a simple interaction, known as “brushing,” so that hovering over a single event in either the text or the image would simultaneously highlight both elements, as well as the corresponding location on the chart’s key (Stasko 2007). For the “lesson” mode, we augmented the features developed for the other two modes with a more guided experience, akin to the lesson that Peabody described in her textbook, through proceduralized interaction (Bogost 2007). In the lesson, users must read each event, one at a time, translate it into color, and then place the colored square in the appropriate location on the grid. Through enhanced user cues, such as converting the cursor to a pointer as it hovers over the grid, and highlighting empty squares as the user hovers over them, users are guided through a digital version of the interactive lesson that Peabody envisioned in print.



Explore

Mouse over each event to see its position on the grid.

1. Table of the Sixteenth Century

7. 1526. Pamphilo de Narvaez attempts to conquer Florida.
8. 1534. James Cartier discovers River St. Lawrence.
9. 1537. Ferdinand de Soto attempts to conquer Florida.
10. 1540. Roberval and Cartier try in vain to colonize Canada.
11. 1542. Ferdinand de Soto dies on the Mississippi.
12. 1562. Coligny sends a Huguenot Colony to America.
13. **1565. St. Augustine, founded by the Spanish Melendez, who conquers the French colony.**
14. 1565. St. Augustine, founded by the Spanish Melendez, who conquers the French colony.
15. 1565. St. Augustine, founded by the Spanish Melendez, who conquers the French colony.

Palette

🟥 England	🟦 Spain	🟩 France
⬜ Unused	⬜ Unused	⬜ Unused

Event Key

1	2	3
4	5	6
7	8	9

1. Battles, Sieges, Beginning of War
2. Conquests, Annexations, Unions
3. Losses and Disasters
4. Falls of States
5. Foundations of States and Revolutions
6. Treaties and Sundries
7. Births
8. Deeds
9. Deaths, of remarkable individuals

Figure 3: A screenshot of the "Explore" mode, with an event from 1565 highlighted. Viewable at http://www.shapeofhistory.net/.

Reimagining Peabody's historical visualization scheme for the web helps to underscore how she understood interpretation as a fundamental part of the process of perceiving visualizations. Her visual design bears very little relation to the immediately intuitive images that we associate with visualization today. And yet, for Peabody, the abstraction of the chart was part of its purpose; she intended her charts to be individually interpreted by each person who encountered them. More than that, she envisioned her charts as lessons in themselves--lessons that often took time and effort in order to complete. In this way, the interactions she envisioned, while made quicker and more intuitive through their digital recreation, lose some of their original intent, in that Peabody did not identify efficiency as a feature of her designs. Instead, she viewed the interpretive process--sometimes difficult and often slow-- as the best source of historical knowledge. The "lesson" of *The Shape of History*, as distinct from Peabody's original scheme, is a reminder of how little interpretation is intended-- even if it is still required--when encountering visualizations of data today.

The Floor Chart: Reimagining Embodiment through Physical Computing

While the digital version of the project emphasizes Peabody's interest in facilitating interaction and interpretation, it does not convey the embodied aspects of the original interaction; looking at a screen is a far different experience than walking around a rug-sized chart on the floor. To reimagine this embodied mode of interaction, we designed a one-meter by one-meter floor chart, consisting of a matrix of thirty by thirty individually addressable light-emitting diodes (LEDs). Each LED corresponds to one subsection of Peabody's original chart, so that the 900 possible events can be represented. (We cannot account for multiple simultaneous events, however). The LEDs can be pre-programmed via custom software, which makes use of Adafruit's NeoPixel library. We are also in the process of developing a flexible touch interface, using conductive copper tape and neoprene, so that the LEDs can be controlled through a soft button-like interaction. Both the LEDs and the touch interface are controlled by an Arduino Mega 2560 microcontroller.

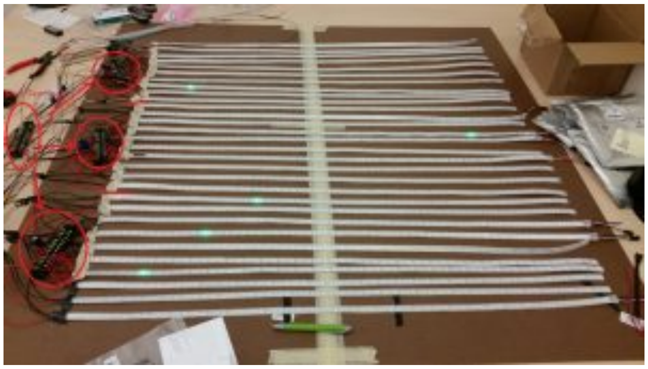


Figure 4. The LED matrix.

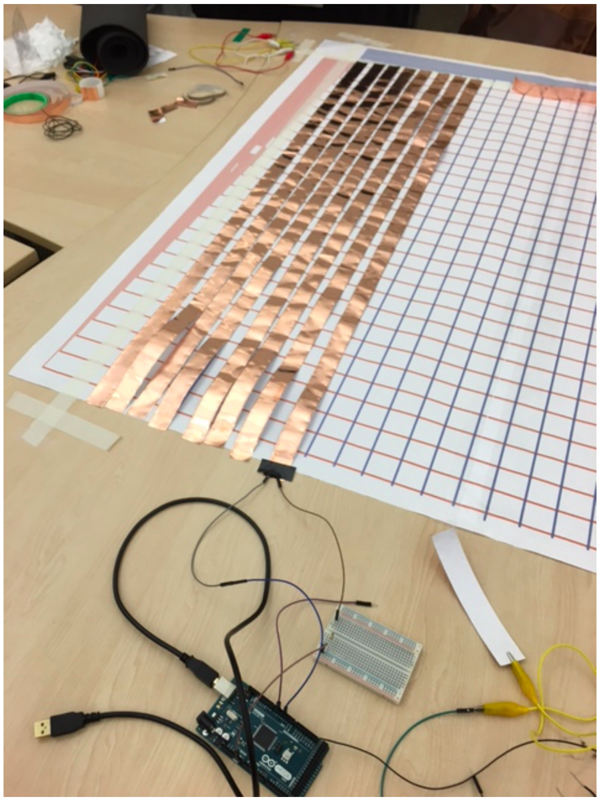


Figure 5: The touch interface in progress.

We view this project as one of speculative design (Dunne and Raby 2013). Since Peabody's original floor charts were not preserved, we must speculate about everything from the size of the chart, to the colors employed, to the events depicted. While we have textual accounts, in Peabody's correspondence, of how nineteenth-century viewers would interact with the floor charts, the original charts were obviously not programmable. What the reimagined floor chart teach us, then, is about how we might incorporate embodied elements into current visualization design practices, as much as about how viewers interacted with large-scale visualizations in the past. It also reminds us about the labor involved in fabricating the original charts. (Peabody complained about the magnitude of the task in her correspondence). The work of data visualization, while not always expressed in physical form, is always the work of many hands.

Conclusions and Next Steps

In their foundational essay on historical fabrication, Devon Elliott et al. observe that “working with actual, physical *stuff* offers the historian new opportunities to explore the interactions of people and things” (2012). In this project, we have sought to extend these opportunities for exploration to include the interactions of people with data, as well as with their visual display. Our project underscores the foundational role of interpretation in designing and perceiving visualizations; and shows how interaction is crucial to the interpretive process. It also points to future modes of visualization, not yet imagined, that might better emphasize embodied ways of knowing. In terms of next steps, for the website, we plan to think through what a more scholarly version of the site, with room for more explanatory text, might look like. For the physicalization, we are continuing to implement the touch interface. From there, we will focus on the aesthetic aspects of the rug, exploring options for light-diffusing fabrics to frame the LEDs, and light-blocking materials to create the grid-lines.

Bibliography

Bogost, I. (2007) *Persuasive Games: The Expressive Power of Videogames*. (2007). MIT Press, Cambridge.

Card, S.K., Mackinlay, J., and Shneiderman, B., eds. (1999). *Readings in Information Visualization: Using Vision to Think*, 1st Edition, Morgan Kaufmann, New York.

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 5(1). Web

Dunne, T and Raby, F. (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT Press, Cambridge.

Elliott, D., MacDougall, R. and Turkel, W. J. (2012). New Old Things: Fabrication, physical computing, and experiment in historical practice, *Canadian Journal of Communication*, 37(1). Web.

Foster, C., Pramer, E., Klein, L. (2016). Repairing William Playfair: Digital Fabrication, Design Theory, and the Long History of Data Visualization. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 513-516.

Peabody, E. (1856). *Chronological history of the United States, arranged with plates on Bem's principle*. Sheldon, Blakeman, New York.

Peabody, E. (1850). *The Polish-American System of Chronology: Reproduced, with some modifications, from General Bem’s Franco-Polish method*. G.P. Putnam, New York.

Posner, M. (2016). What’s Next: The Radical, Unrealized Potential of Digital Humanities. In *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein. University of Minnesota Press, Minneapolis. Web.

Rhoda, B. (1999). *Elizabeth Palmer Peabody: A Reformer on her Own Terms*. Harvard University Press, Cambridge.

Sayers, J. (2015). Why Fabricate?, *Scholarly and Research Communication*, 6(3), n.p.

White, A. (2011). *The Elements of Graphic Design, 2nd ed*. Allworth Press, New York.

Yi, J.S., Kang, Y., Stasko, J.T., and Jacko, J.A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics* 13 (6), pp. 1224-1231.

Integrating Humanities and Science: the Scriptospatial Visualization Interface

Fenella G. France
frfr@loc.gov
Preservation Research and Testing Division
Library of Congress, United States of America

Alberto Campagnolo
acamp@loc.gov
Preservation Research and Testing Division
Library of Congress, United States of America

Introduction

Advances in technology and digital access have paved the way for the improved utilization and interpretation of scientific analyses of humanities materials for digital humanities studies. Integrating scientific analyses with humanities and curatorial knowledge (STEM: science, technology, engineering, and math, to STEAM: science, technology, engineering, **art**, and math) is a critical multidisciplinary approach for expanding the full potential of scientific techniques and technological advances, and realigning complementary disciplines that have been artificially segregated. Scientists and curators have exposed hidden and previously unknown contextual information within original source materials, such as changing “subjects to citizens” on the Rough Draft of the Declaration of Independence (Library of Congress, 2010). Hyperspectral imaging provides additional data layers by capturing images of documents in distinct narrow waveband regions of the visible and non-visible spectrum—from ultraviolet through visible to infrared. The cube of captured digital image files contains a wealth of information, but requires significant interpretation to process and analyze the collected data. Scientists, scholars, and students in both art and science disciplines have been collaborating to glean new information from historical manuscripts. The Library spectral imaging program includes a spectral reference database and integration of data from other non-invasive analytical techniques to create a full analytical mapping of heritage documents and objects for non-destructive analyses of collection materials (France, 2016)

Digital imaging capabilities allow researchers to characterize pigments and inks on the document, retrieve hidden and lost text, and illuminate production and creation methods. The range of data captured allows greater access to the information available from fragile historic documents, including the 1507 Waldseemüller World Map and the 1513 Ptolemy *Geographia*, where investigations revealed links to the same original printing location (France, 2016). *Scriptospatial* (a term originally coined by Toth and France to refer to the viewing of associated imaging and materiality data linked on an image of a historic document) refers to applying a spatial information system approach to documents, creating an interactive interface for scholars and scientists to interact with the object and the data. Scriptospatial representations of digital data from documents utilize an accurate coordinate system that links scientific and scholarly analyses to the creation of a new digital cultural object (DCO), allowing inferences to be drawn to generate new knowledge. This approach to viewing digital cultural materials in multiple layers applies an archaeological approach toward uncovering and interconnecting information strata of historic and modern documents. Scriptospatial mapping of documents with an accurate coordinate system allows the layering of scientific and scholarly analyses to the DCO. This allows inferences to be drawn to generate new knowledge through analysis of the data linked to spatial points (or areas). This approach to viewing the DCO applies a GIS methodology toward uncovering and interconnecting information layers of cultural heritage artefacts, just as in the case of archaeological strata. Utilizing an object-oriented approach in conjunction with the spatial data layers allows the mapping of spatial and temporal data with increasing complexity. Examining and explaining the physical, spectral and chemical properties of these historic materials permit scientists and scholars to link these scientific analyses to other data about the creation of the object.

Digital spectral imaging of cultural heritage objects at the Library of Congress has capitalized on over a decade of research and development into not only spectral imaging and processing, but also the development of standard spectral image products (France et al, 2010). Advances over the past decade by an experienced team have led to an advanced capability to study cultural heritage objects with a robust spectral imaging system that provides large-format, high quality images and standardized data output using advanced commercial off-the-shelf components (Christens-Barry et al, 2009).

Developing an object-oriented approach to data access and sharing requires integration of spectral imaging with data from other sources in a variety of formats. This requires effective spatial metadata to allow linkages to specific locations within the images. This is necessary not only to register locations on the same section of a manuscript leaf in various spectral bands, but also to link other images and transcriptions with the spectral images. Based on geospatial mapping and layering of data used to identify points on satellite images, the same technologies, work processes and skills can be applied to spectral images of manuscripts: A camera collecting images over a manuscript is similar to a satellite collecting geospatial data over the Earth. Using technologies developed for "geospatial" systems to link each point on the globe with images from earth resource satellites and data collected from other instruments, spectral imaging can link the "scriptospatial data" from each point on a manuscript with images from various imaging and scientific devices. This provides a standardized method to support links between images and data from the same object location.

With multiple data entries for samples, precisely defining the specific point where the sample or scientific data collection takes place is critical in comparing data from different research types or objects. For samples taken from a larger, non-uniform, heterogeneous object such as a manuscript, textile or painting, the spatial location of the sample point on the object must be defined to be able to integrate the data from various research tools. Spatial metadata elements will allow linkages to specific locations on an object, potentially within images of the objects. Scriptospatial data can serve as an interface for scientific dialogue in "one shared layer," linking data from various sources for in-depth studies and analyses of a specific research topic or object.

In many current research databases, the metadata elements for spatial location are not provided to capture detailed data on where an instrument collects data, or a sample is taken. This is not an issue for uniform, homogeneous samples of paints, pigments, media or other samples, but is critical for samples taken from a heterogeneous object like a painting, manuscript or textile. By defining a Cartesian coordinate system on an object or image of an object, as well as the degree of precision required, specific sample points on an object can be defined. This allows integration with other images of the same object and scientific samples from the same point.

In its Scriptospatial Visualization Initiative, the Library of Congress PRTD has capitalized on developments with geospatial systems to apply Thermopylae "i-spatial" support and Google Map tiling and data formatting to the integration of large and complex visual scriptospatial datasets populated with scientific data from various instruments, research topics or objects. This provides data access in "one shared layer" of scientific data. This is an important first step in capitalizing on the three decades of technology development by the GIS community to advance preservation science and cultural heritage data sharing and research. The additional unique component will be the layering of scholarly research interpretations and publications, enabling ease of access to a rich resource of data directly linked to the original object. This will reduce the challenges faced with searching through data that is not well catalogued, or yet searchable without this expanded document interpretation and linking of scholarly knowledge.

Integrated access to associated source material scientific data adds contextual value, provenance information, and can reveal non-visible information hidden within the original document. The Scriptospatial approach allows layering of scientific and scholarly or curatorial research data within one location, enhancing the analysis and depth of knowledge off the original document, and creating a more effective interface for interaction with historic materials (France et al, 2010). The innate layering of multiple sources of information of the Shared Canvas Model (Sanderson and Albritton, eds, 2013) onto one view of an object, coupled with the strong annotation capabilities of the Web Annotation Data Model (Sanderson et al, eds, 2016) raises interest in the possible integration of the Scriptospatial model with IIIF efforts (International Image Operability Framework). One of the challenges we face with the current data deluge, is how to effectively access, select and link appropriate data and information. Scriptospatial is a value-added data approach, creating cohesive structured management of multidisciplinary data. The authors have been engaging with other US, European and United Kingdom colleagues to create a cohesive integrated and collaborative approach to this visualization.

Bibliography

Christens-Barry, W. A., Boydston, K., France, F. G., Knox, K. T., Easton, R. L. Jr., and Toth, M. B. (2009). "Camera System for Multispectral Imaging of Documents." In *SPIE Proceedings Vol. 7249: Sensors, Cameras, and Systems for Industrial/Scientific Applications X*, edited by Erik Bodegom and Valérie Nguyen, 7249:8–10. San Jose (CA). doi:10.1117/12.815374.

France, F. G. (2016). "Spectral Imaging: Capturing and Retrieving Information You Didn't Know Your Library Collections Contained." In *What Do We Lose When We Lose a Library? Proceedings of the Conference Held at the KU Leuven* 9-11 September 2015, edited by Lieve Watteeuw and Mel Collier, 189–97. Leuven: KU Leuven University Library.

France, F.G., Emery, D., and Toth, M. B. (2010). "The Convergence of Information Technology, Data, and Management in a Library Imaging Program." *Library Quar-terly.* Special Edition: Digital Convergence: Libraries, Archives, and Museums in the Information Age 80 (1): 33–59.

International Image Interoperability Framework. (2016) "IIIF | International Image Interoperability Framework." http://iiif.io/ (accessed 01-11-2016).

Library of Congress. (2010). "Hyperspectral Imaging by Library of Congress Reveals Change Made by Thomas Jefferson in Original Declaration of Independence Draft." *Press Release*, July 2. https://www.loc.gov/today/pr/2010/10-161.html (accessed 01-11-2016).

Sanderson, R., and Albritton, B., eds. (2013). "Shared Canvas Data Model 1.0." http://iiif.io/model/shared-canvas/1.0/ (accessed 01-11-2016).

Sanderson, R., Ciccarese, P., and Young, B, eds. (2016). "Web Annotation Data Model." https://www.w3.org/TR/annotation-model/ (accessed 01-11-2016).

Reading the Norton Anthologies: Databases, Canons, and "Careers"

Erik Fredner
fredner@stanford.edu
Stanford University, United States of America

David McClure
dclure@stanford.edu
Stanford University, United States of America

JD Porter
jdporter@stanford.edu
Stanford University, United States of America

What type of canon do the Norton anthologies of literature construct? And how has that canon changed over time? These questions are somewhat unusual for humanists in that their answers could be framed not in a syllogism or thesis, but rather in the forms of the list and the table—and extensive ones at that.

We wanted, first, a way to see who goes in and who goes out of this canon that we so often teach from. In a sense, this project is as much about pedagogy as it is about literary criticism: We have no investment in the Nortons as being representative of "The Canon," but rather see them as a medium through which undergraduate and graduate students of literature encounter major works and begin to formulate their ideas about the literary field. Following John Guillory, the Nortons seem to be one of the primary means by which the cultural capital of literature gets distributed and reinscribed within the university. How, then, has this medium—of texts and canons that inform courses, students, and scholars—changed along with literary criticism over the past half-century?

Our team built a database containing every work and excerpt featured in the Norton Anthologies that we have studied so far, with room to grow for those that remain. This allows us to easily see what we have been thinking of as the "careers" of both authors and individual works over time. How, for instance, have the works selected to represent Milton changed over time? When was Margaret Atwood first added to any Norton Anthology? Which poems represent Langston Hughes in the *Anthology of Poetry*? Are they different from those that represent him in the anthologies of *World*, *American*, or *African-American* literature? What proportion of authors in the anthologies are women, and how has that changed over the last fifty years? Which authors have been cut from the anthologies? And which authors or works replaced them?

In order to answer these questions about the people and ideas admitted to these canons, we needed to restructure the data from the Nortons' tables of contents into a format that could be queried and would reveal the relationships among many different works and selections from works across a variety of different manifestations. The problems this poses from the perspective of data structure are easiest to think through with a major author like Shakespeare, who appears in every anthology relevant to his work. We need to know which works were selected to represent Shakespeare in each edition of every anthology in which his works appear. For example, which Shakespeare plays appeared in the first edition of the *Norton Anthology of English Literature*, and which in next eight? How do those selections compare to the ways in which Shakespeare is represented in the anthologies of *Drama*, *Poetry*, *Western*, and *World* literature, across each of their individual editions?

We achieved this by creating a structure based on a set of *n*-deep parent-child relationships and a number of many-to-many connections, using a web interface for parallel data entry and validation between several collaborators simultaneously. Using this structure and the Shakespeare example above, *King Lear* becomes a "child" of Shakespeare, and Lear's "Blow, winds … !" speech from Act III a child of *King Lear*. Because of this nesting, we can then measure not only which anthologies *any* work of Shakespeare's appears in, but, of those, which contain *King Lear* in full, which only have the excerpt of Lear's speech, and which contain other parts of *Lear*. This allows us to be more precise about the ways in which we count authors' presence and absence across all of the anthologies that this project will eventually consider. This data entry interface was built with the Django, an open-source web application framework, and the code is publicly accessible on GitHub. The database will be demonstrated and described in more detail during the presentation.

Having produced new editions and types of anthologies semi-regularly for more than fifty years, W.W. Norton & Company has been in the business of binding literary canons longer than anyone else still publishing. Since M.H. Abrams edited the first *Norton Anthology of English Literature* in 1964, numerous editions and kinds of anthologies have followed: *The Norton Anthology of American Literature*, *World Literature*, *Western Literature*, *Poetry*, *Drama*, *Theory and Criticism*, *Short Fiction*, *Literature by Women*, *African American Literature*, *Latino Literature*, *Jewish American Literature*, etc.

Much can be learned about the ways in which the Nortons were designed from these titles alone. First, the largest anthologies are defined by both geography and a linear temporality influenced by conventions of periodization. *The Norton Anthology of World Literature* lays claim to it all, from Afghanistan to Zimbabwe, and from *Gilgamesh* to Orhan Pamuk. *Western Literature* claims a smaller (if vaguer) part of the world, and *English* and *American* literature focus on national literatures, including postcolonial and expatriate writers within the bounds of the nation-concept. While these geographic anthologies are ostensibly genre-agnostic, others are genre-specific (*Poetry*, *Drama*, *Short Fiction*). And the last type focuses on writing by and about writers of a specific gender (*By Women*), ethnicity (*Latino*), or religion (*Jewish*).

One of the premises we read as implicit in the Norton's design, then, is that some authors and works become significant enough to include only in specific contexts. Making it into the *World Literature* anthology seems to denote significance at a greater level than inclusion in the *Western Literature* anthology alone would. Likewise, seeing a writer anthologized in *Short Fiction* but not *World Literature* seems to imply a significance limited to that literary form. Canon formation has always relied on a logic of a ranking or tiering, and the pool of authors and texts against which a given work "competes" is greatest at the largest scale of population. We argue that the geographically bound Nortons can be read in such a way that they imply a hierarchy even among the canons the anthologies already connote.

Of course, all of the decisions we measure across these many tables of contents are underwritten by a human element. Many practical and historical factors that exist at a slant to the question of a work's "canonicity" attend the production of an anthology that stretches to more than 6,000 pages, serves tens of thousands of instructors and students, and our analysis attempts to account for these factors. A quantitative approach is necessary but not sufficient to read the ways in which the Nortons have represented and continue to represent works that, taken together, lay claim to the status of a national, generic, or global literature. Two key examples of the incommensurability of the form to its implicit claims: Because of its length, the novel is poorly served by the anthology form. Some shorter novels and novellas do get anthologized. But, more often than not, writers who are primarily known as novelists are represented by a single short story, or an excerpt from a novel. The second overarching practicality is the influence of authorial estates and the cost of printing rights, especially for 20th and 21st century authors.

As a way of approaching these institutional and qualitative questions, we have begun a set of interviews with Martin Puchner, the current general editor of the *World* and *Western* anthologies, and will discuss some of his insights into the decision-making process in the presentation. Among these include the influence of instructor and student surveys on texts that get selected, the impact of rights costs on the texts that get chosen for a given author, and the place of editorial intervention in relation to these powerful practicalities.

Our database and the attendant institutional research on the Nortons as the product of both scholarly editing and the demands of the market allow us to see the trajectories not only of individual authors and works, but broader trends of inclusion and exclusion in the Norton's canon. By gathering data about both the works and the authors who wrote them, we reveal the ways in which the Norton has responded to the expansion of the literary canon, growing in size while simultaneously giving a greater share of its pages to authors and ideas that would not have been considered canonical 1962. In the process, we find authors whose literary reputation has waxed or waned (or both); those whose names have been a constant presence, but whose representative works have dramatically changed; those who were slated for canonization but never "made it;" and those who have arrived late but seem to be here to stay.

Like many of the Stanford Literary Lab's projects, "Reading Norton Anthologies" operates at several scales at once. We are interested both in individual texts and authors, as well as broader patterns of representation and contextualization within the confines of this object that occupies liminal spaces between statement and syllabus, and between the market and the canon.

Bibliography

Bausch, R., ed. (2015). *The Norton Anthology of Short Fiction*. Eighth. W.W. Norton & Company.

Baym, N., ed. (2012). *The Norton Anthology of American Literature*. Eighth. W.W. Norton & Company.

Bourdieu, P. (1993). *The Field of Cultural Production*. Columbia University Press.

Ferguson, M., Salter, M. J., and Stallworthy, J., eds. (2005) *The Norton Anthology of Poetry*. Fifth. W.W. Norton & Company.

Gates Jr, H. L. (2014) *The Norton Anthology of African American Literature*. Third. W.W. Norton & Company.

Gilbert, S., and Gubar, S., eds. (2007) *The Norton Anthology of Literature by Women: The Tradition in English*. W.W. Norton & Company.

Greenblatt, S., ed. (2012) *The Norton Anthology of English Literature*. Ninth. W.W. Norton & Company,

Guillory, J. (1995). *Cultrural Capital: The Problem of Literary Canon Formation*. University of Chicago Press,

Puchner, M., ed.(2014) *The Norton Anthology of Western Literature*. 9th ed. W.W. Norton & Company.

Puchner, M., ed. (2012). *The Norton Antholog of World Literature*. 3rd ed. W.W. Norton & Company.

Shakespeare, W. (2005). *King Lear*. Edited by George Hunter. Penguin Classics. Penguin.

Stavans, I., ed. (2011) *The Norton Anthology of Latino Literature*. First. W.W. Norton & Company.

The Intellectual Structure of Digital Humanities: An Author Co-Citation Analysis

Jin Gao
jin.gao.13@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Oliver Duke-Williams
o.duke-williams@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Simon Mahony
s.mahony@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Melanie Ramdarshan Bold
m.bold@ucl.ac.uk
University College London, United Kingdom

Julianne Nyhan
j.nyhan@ucl.ac.uk
UCL Centre for Digital Humanities, United Kingdom

Introduction

With many ongoing debates (Gold, 2012) and "unwritten" histories (Nyhan and Flinn, 2016), the research practice of the Digital Humanities (DH) has been around for 70 years. Many works have been trying to draw general conclusions of the disciplinary structure (McCarty, 2003; Gold, 2012; Terras et al., 2013; Schreibman et al., 2016; Nyhan and Flinn, 2016), and have pointed to the potential usefulness to analyse the discipline from statistical aspects. The usefulness focuses on describing the intellectual structure, scholarly interactions and disciplinary development of DH. Some studies have dedicated their attention to these matters (Grandjean, 2016; Nyhan and Duke-Williams, 2014; Quan-Haase et al., 2015; Wang and Inaba, 2009), or have focussed on one of these topics (Sugimoto et al., 2013), but few of them have engaged either with the bibliometric network method, or with the latest large-scale scholarly datasets to study the DH community as a whole.

Therefore, to fill this gap, based on a provisional dataset that has been compiled from core DH journals, this study performs an exclusive all-author co-citation analysis (ACA) with the 200 most cited scholars by fractional citation count to map and demonstrate the intellectual structure and to identify the most influential scholar groups and topics within DH.

To the best of our knowledge, this study is the first to apply bibliometric methods to visualise DH knowledge structure and the scholar clusters. This research output will make a valuable contribution to the current discussions and debates about DH knowledge structure and wider scholarly networks.

Methodology

With ACA as the main methodology, the research contains four steps, and each with a different methodology: building a DH citation index according to the publications of these journals; selecting authors as the core objects for citation analysis; assigning scholars to different distance-based clusters by calculating the author co-citation matrix to similarity matrix (Waltman and van Eck, 2013); finally, visualising the DH citation network which aims to show the scholar clusters, and the knowledge structure and diffusion of DH.

The three DH core journals that our dataset has been constructed from are: "*Computers and the Humanities*" (*CHum*), "*Digital Humanities Quarterly*" (*DHQ*), "*Literary and Linguistic Computing*" (*LLC*) (now "*Digital Scholarship in the Humanities*") (*DSH*). The bibliographies as well as the metadata of all their publications (including the reviews and editorials etc.) published until June 2016 have been collected. It should be noted that none of these journals spanned the whole period selected (1966-2016): *CHum*, the first DH journal started in 1996, and ceased publication in 2004; *LLC/DSH* began in 1986; *DHQ* began in 2007. Figure 1 shows the total publications each year from 1966 until June 2016 for these journals.

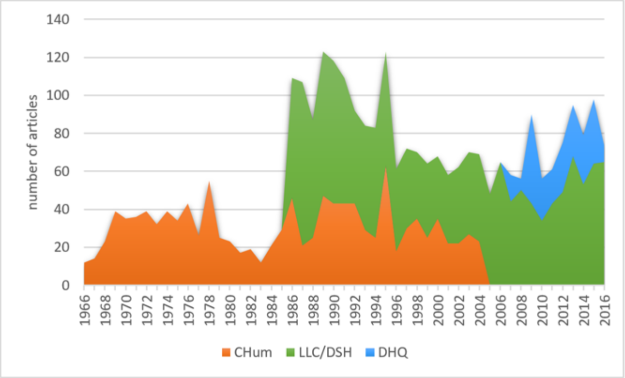


Figure 1. In total, 3,068 journal articles: *CHum* (1,195 articles with 26,033 citations), *LLC/DSH* (1,633 articles with 28,501 citations), and *DHQ* (240 articles with 4,289 citations)

Author co-citation analysis (ACA) can reveal the intellectual structure of a field from its academic publications by calculating the frequencies with which two authors are cited together. That is to say, if an article cites at least one article of author A, and at least one of author B that is different from the one of A, the co-citation count increases by 1. The more co-citations two authors receive, the more likely their publications and researches are related (Bellardo, 1980). Therefore, the clusters of related authors indicate the networks of research topics, or influential focuses within a discipline.

The initial findings with the top cited 200 authors displayed on the maps (see the provisional maps in Figure 2 and Figure 3) have provisionally revealed five sub-fields within DH.

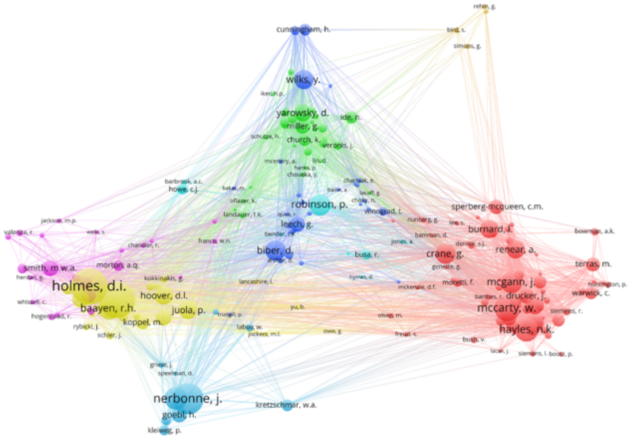


Figure 2. The provisional ACA network map in DH, data from journals *CHum*, *LLC/DSH*, and *DHQ*, 1966-2016, created using VOSviewer

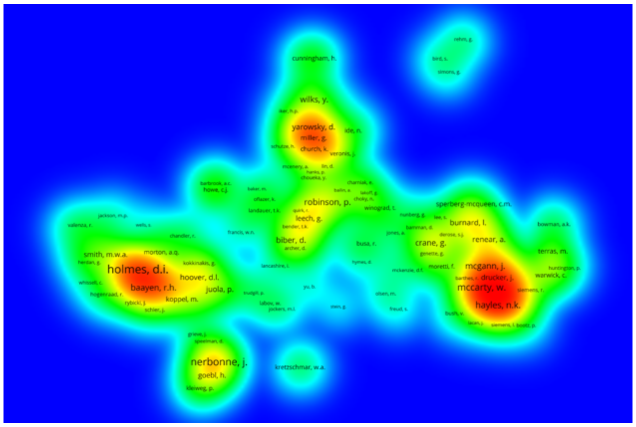


Figure 3. The provisional ACA density map in DH, data from journals *CHum*, *LLC/DSH*, and *DHQ*, 1966-2016, created using VOSviewer

Both of the maps (Figure 2 and Figure 3) are distance-based. Each node on the map represents an author, and the distance between two authors is their relations (the closer the distance, the stronger the connection). Authors are distributed quite unevenly, and this makes it easy to identify clusters of related nodes. The size of the node represents the citation count this author received, and the higher the citation count is, the bigger the node. On the density map, the density value depends on the size, number and distance of the nodes around it, so the higher the density value, the colour is more red than blue.

Both maps have revealed the general structure of the scholarly communication between DH scholars via publications. Horizontally across the centre of the map, there is a loosely connected circle of five DH scholar clusters: centre (focused on "Leech, G"), top (focused on "Miller, G"), bottom (focused on "Nerbonne, J"), left (focused on "Holmes, D.I"), and right (focused on "McCarty, W"). The clusters distribution on the density map reveals that there is a clear separation between top, centre, right clusters to left and bottom clusters. Especially the right cluster (focused on "McCarty, W") and the left cluster (focused on "Holmes, D.I") turn out to be denser than other clusters. This shows that these two clusters are more significant and have more citation influence. According to the provisional analysis, these five clusters appear to be associated with five different DH research topics: English study at the centre; general historical literacy and information science on the right; language modelling and natural language processing at the top; statistics and text analysis on the left; computational linguistics particularly on Dutch and German speaking at the bottom. These five clusters, however, are also grouped into two different bigger groups. The English study, language modelling, and general historical literacy seem to be in one group which is more related, while the statistics and Dutch-German linguistics are also very closely related to each other.

Limitations and Future study

This research is part of the first author's ongoing PhD study, funded by UCL ORS scholarship and based at the UCL Centre for Digital Humanities. The doctoral reseach maps DH intellectual, social and environmental structures using the Invisible College model (Zuccala, 2006).

There are some limitations that need to be noted, such as the citation lag time. In order to build up a citation record for co-citation, it takes around five to eight years (Hopcroft et al., 2004). This could explain that certain recognisable authors might not appear on the maps yet. Also, because the co-citation method studies the knowledge base as its subject, the map emphases more on authors published some time ago, which might not include the "new comers".

In the future work, the ACA study will be extended to include more citation data. The ACA study will be divided into discreet periods to construct maps of different DH development stages. Given that different journals have different topical foci, the research will also analyse individual journal to discover its attribute.

Bibliography

Bellardo, T. (1980). "The use of co-citations to study science." *Library Research*, 2(3): 231–237.

Gold, M. K. (2012). *Debates in the Digital Humanities*. University of Minnesota Press.

Grandjean, M. (2016). "A social network analysis of Twitter: Mapping the digital humanities community." *Cogent Arts and Humanities*, 3(1): 1171458.

Hopcroft, J., Khan, O., Kulis, B., Selman, B., (2004). "Tracking evolving communities in large linked networks." *Proceedings of the National Academy of Sciences*, 101: 5249–5253.

McCarty, W. (2003). "Humanities Computing." In *Encyclopedia of Library and Information Science*. New York: Marcel Dekker.

Nyhan, J., and Flinn, A. (2016). *Computation and the Humanities: Towards an Oral History of Digital Humanities*. Cham: Springer International Publishing.

Nyhan, J., and Duke-Williams, O. (2014). "Joint and multi-authored publication patterns in the Digital Humanities." *Literary and Linguistic Computing*, 29 (3): 387-399.

Quan-Haase, A., Martin, K., and McCay-Peet, L. (2015). "Networks of Digital Humanities Scholars: The Informational and Social Uses and Gratifications of Twitter." *Big Data and Society*, 2(1).

Schreibman, S., Siemens, R. G., and Unsworth, J. (Eds.). (2016). *A New Companion to Digital Humanities*. Chichester, West Sussex, UK: John Wiley and Sons Inc.

Sugimoto, C. R., Thelwall, M., Larivière, V., Ding, Y., and Milojević, S. (2013). *Mapping Digital Humanities*. Retrieved from http://did.ils.indiana.edu/dh/

Terras, M., Nyhan, J., and Vanhoutte, E. (Eds.). (2013). *Defining Digital Humanities: A Reader*. Farnham, Surrey, England : Burlington, VT: Ashgate Publishing Limited ; Ashgate Publishing Company.

Waltman, L., and van Eck, N. J. (2013). "A smart local moving algorithm for large-scale modularity-based community detection." *The European Physical Journal B*, 86(11).

Wang, X., and Inaba, M. (2009). "Analyzing Structures and Evolution of Digital Humanities Based on Correspondence Analysis and Co-word Analysis." *Art Research*, 9: 123–134.

Zuccala, A., (2006). Modeling the Invisible College. Journal of the Association for Information Science and Technology. 57: 152–168.

Negotiating Sustainability: The Grant Services "Menu" at UVic Libraries

Lisa Goddard
lgoddard@uvic.ca
University of Victoria, Canada

Christine Walde
cwalde@uvic.ca
University of Victoria, Canada

Brief summary

This paper provides a brief overview of library best practices for digital curation, with particular attention to the areas that highlight disciplinary tensions between library science and the humanities. The authors introduce the University of Victoria's grant service "menu" for digital preservation and hosting services, and outline some of the most promising models for balancing creativity with sustainability in DH project design. We will suggest roles for libraries, researchers, administrators, and funders in helping to create technical and social conditions that nurture sustainable research projects in the digital humanities and beyond.

Abstract

Knowledge building is an iterative process that refines and extends previous research. Through citation, we acknowledge our debt to scholars and theorists whose work enables our own. The ephemeral nature of the digital world threatens to destabilize a centuries long system of scholarly communication and knowledge sharing. In a print ecosystem many immutable copies of an object are distributed globally and are curated by network of organizations. In the digital world a single copy of an object is served from a central location. Digital content is thus susceptible to manipulation, corruption, and erasure. The key to analog preservation is to ensure that artefacts remain the same. Digital preservation, in contrast, requires "active management" comprising constant changes, patches, and updates. Objects become quickly obsolete as the environments around them change.

Funding agencies are putting increased pressure on researchers to include sustainability plans in funding applications (NEH 2016, SSHRC 2016). Researchers often turn to the University Library to provide preservation solutions for digital projects without fully understanding the technical, policy, and funding implications of these requests. Libraries have made significant strides in planning for the long-term preservation of the many thousands of digital objects in our collections. Digitization projects adhere to strict standards for resolution, colour management, and file formats (FADGI 2016). Digital asset management systems like Hydra/Fedora provide a single place to store objects along with descriptive and administrative metadata that helps to determine the preservation actions that should be taken against each object (Goddard, 2016). Those actions include auditing and bit-checking of file systems to ensure against data loss, format migrations as media and file types become obsolete, replication of objects across different technology stacks and jurisdictions, and discovery interfaces that ensure continued discoverability and access. Libraries are building national networks that will allow us to replicate data across multiple jurisdictions to mitigate against disasters both natural and human (DPN 2016, Canadiana 2016, CARL 2016). Despite concerted efforts, only a handful of library repositories have so far met the stringent conditions that are necessary for certification as a Trusted Digital Repository, which requires technical and policy elements including plans for long term staffing and funding, and contingency plans in the event of organizational failure (CRL, 2015). Ultimately, libraries still can't make guarantees about preservation for digital objects in our own collections, even those that are subject to internationally recognized best practices. This problem is compounded when DH research projects fail to adhere to adequate quality standards for objects (e.g. images, texts, video, maps, mark-up) and overlook established metadata models and vocabularies.

To this point we have outlined the challenges of curating fairly static digital files, but most DH projects are far more than the sum of their digital objects. Many DH research projects are complex software stacks with many layers of tools, objects, code and dependencies. If a project is built on Drupal, for example, librarians will have to not only maintain all of the unique objects and code produced by the project, but they are also committed to maintaining a specific version of a rapidly evolving software platform -- a version that will likely be obsolete before the project concludes. Drupal is, at the very least, well documented and widely deployed. Many DH projects also include custom-built tools, the inner workings of which are known only to a handful of people on the research team. The complex technology profiles of contemporary DH projects require ongoing active management including patching, tending, and rebuilding over time (Burpee, 2015). While the library may have sufficient resources to steward one or two unique project environments, this approach cannot scale to hundreds or even thousands of projects over time. In the current technical and funding environment is simply not possible for libraries to provide high-level curation for the enormous variety of funded digital projects that are produced by researchers within their organizations.

Libraries alone will not solve the problem of sustainability in DH projects. A fundamental characteristic of sustainability is that it must be established as a key design principle from the outset. It is almost impossible to retroactively render a project sustainable without rebuilding from the ground up. Initial choices about technologies, data models, formats, and documentation will influence the likelihood that a project will still be accessible in a decade. One complicating factor is that sustainability is largely at odds with a researcher's freedom of choice when it comes to decisions about platforms, tools, and data models. Truly sustainable DH projects will require a level of standardization that is far from the current norm in DH project development, and which is unlikely to be unequivocally embraced by humanists. Research is an experimental process, and technological constraints can stifle creativity and independence of action. Models, by their very nature, seek to simplify, while the humanist tradition revels in nuance and complexity (McCarty, 2005; Quamen, 2013).

Leslie Johnston from the Library of Congress suggests that libraries can pursue two models for preserving complex DH projects. The first approach is to "preserve the content but forgo the look and feel. This is often extremely unpopular." The second is to "preserve the content and the look and feel exactly as they were implemented. This is often close to impossible." (Johnson, 2013) The tension between these two models is where libraries, researchers, and funders need to more clearly outline our assumptions and expectations.

The University of Victoria Libraries have developed a suite of preservation services for grant funded projects in order to plainly articulate our competencies, assets, and constraints (Goddard and Walde, 2017). This document acts as kind of a "menu" of services from which researchers can select as they develop their grant applications. These include the use of our Hydra/Fedora4 digital asset management system, metadata expertise that extends to consultations around interoperability and linked data, web hosting and discovery, exhibit building software, copyright consultation, open access publishing, research data management, and digitization services. We provide template paragraphs related to sustainability, preservation, open access, and knowledge mobilization that researchers can easily repurpose for any given funding proposal. We include a break down of the in-kind value of each of these services, along with any costs that will be charged back, so that researchers can easily estimate the value of the institutional commitment. We hope that this approach will enable critical conversations about sustainability to happen during the grant writing process, rather than towards the end of funding cycles as has been too often the case in the past. In order to offer our "gold standard" preservation services libraries will have to be involved in early conversations about technology preferences and data models. We certainly don't assume that all decisions will be dictated by curation needs, but rather that our consultation will enable researchers to make clear-eyed decisions about the impact of their choices on sustainability.

The preservation "menu" is an appealing model for researchers, as it enables them to quickly understand the variety of services and in-kind contributions that the library can offer in order to strengthen a funding application. There are also advantages to the library. By tying preservation services to grant funded projects we can avail of a rigorous review process that helps us to direct library resources towards research that has been deemed valuable by a network of disciplinary experts. It provides an easy formula for calculating in-kind contributions for letters of support, and to some extent standardizes the process of writing those letters. It helps to promote librarians as desirable co-applicants and collaborators on funding applications. It underscores to administrators the library's value as a university research support. This model also provides mechanisms whereby grant funds can flow back into the development of new features for the library's digital asset management and publishing platforms.

Susan Brown notes that "successful technologies rely on social resources." (Brown, 2016) Part of our challenge is to muster support from researchers, librarians, administrators, and funders to create optimal conditions for long-term digital curation. The conversation about long-term preservation will be an ongoing negotiation that bridges different disciplinary perspectives, and balances ideals with resource constraints. Just as the traditional model of scholarly print publishing has shaped the means of scholarly production through the last two centuries, these conversations will ultimately will help to shape the future of humanities research platforms, resources, and methodologies.

Bibliography

Brown, S. (2016). "Tensions and tenets of socialized scholarship." Digital Scholarship in the Humanities, 31(2): 283–300. http://doi.org/10.1093/llc/fqu063

Burpee, K. J. (2015). "Outside the Four Corners: Exploring Non-Traditional Scholarly Communication." Scholarly and Research Communication, 6(2). http://src-online.ca/index.php/src/article/view/224/417.

Canadian Association of Research Libraries (CARL) (2016). Compute Canada and the Canadian Association of Research Libraries Join Forces to Build a National Research Data Platform. Ottawa, Ontario: CARL. https://www.computecanada.ca/research/compute-canada-and-the-canadian-association-of-research-libraries-join-forces-to-build-a-national-research-data-platform/

Canadiana (2016). Preservation Policy and Strategy. Ottawa, Ontario. http://www.canadiana.ca/preservation-policy-strategy

Center for Research Libraries (CRL) (2015). Certification and Assessment of Digital Repositories. Chicago, IL. https://www.crl.edu/archiving-preservation/digital-archives/certification-assessment

Digital Preservation Network (DPN) (2016). Node Requirements. http://dpn.org/dpn-admin/resources/dpnnodereqmarch2016.pdf

Federal Agencies Digitization Guidelines Initiative Working Group (FADGI) (2016). FADGI Technical Guidelines for Digitizing Cultural Heritage Materials. Washington, DC. http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image_Tech_Guidelines_2016.pdf

Goddard, L. (2016). "The Read-Write Library." Scholarly and Research Communication. 7(2). http://dx.doi.org/10.22230/src.2016v7n2/3a255

Goddard, L. and Walde, C. (2017). Hosting and Preservation Services for Grant-Funded Research Projects. Victoria, BC: University of Victoria. http://www.uvic.ca/library/about/ul/UVicLibraries_GrantServices_Feb2017.pdf

Johnston, L. (2013). "Digital Humanities and Digital Preservation." The Signal. https://blogs.loc.gov/thesignal/2013/04/digital-humanities-and-digital-preservation/

Kretzschmar, W. and Potter, W. (2010). "Library collaboration with large digital humanities projects." Literary and Linguistics Computing, 25(4): 439–45. http://dx.doi.org/10.1093/llc/fqq022

Marcum, D. (2016). "Due diligence and stewardship in a time of change and uncertainty." Ithaka S+R Issue Brief. https://doi.org/10.18665/sr.278232

McCarty, W. (2005). "Chapter 1: Modelling." Humanities Computing. London, UK: Palgrave, pp. 20-72.

Muñoz, T. and Flanders, J. (2014). "An Introduction to Humanities Data Curation." Digital Humanities Data Curation Guide. http://guide.dhcuration.org/contents/intro/

National Endowment for the Humanities (NEH) (2017). Data Management Plans for NEH Office of Digital Humanities Proposals and Awards. https://www.neh.gov/files/grants/data_management_plans_2017.pdf

Quamen, H. (2013). "The Limits of Modelling: Data Culture and the Humanities." Scholarly and Research Communication, 3(4). http://src-online.ca/index.php/src/article/view/69.

Social Sciences and Humanities Research Council of Canada (SSHRC) (2016). Tri-Agency Statement of Principles on Digital Data Management. Ottawa, Ontario. http://www.science.gc.ca/default.asp?lang=En&n=83F7624E-1

Vandegrift, M. and Varner, S. (2013). "Evolving in Common: Creating Mutually Supportive Relationships Between Libraries and the Digital Humanities." Journal of Library Administration, 53(1): 67–78.

"A Trace of this Journey": Citations of Digitised Newspapers in UK History PhD Theses

Paul Matthew Gooding
p.gooding@uea.ac.uk
University of East Anglia, United Kingdom

> "In two weeks, despite these notes, I shall no longer believe in what I am experiencing now. One must leave behind a trace of this journey which memory forgets" (Cocteau, 2013).

Academic citations are prostheses for the scholarly memory, providing traces of a text's origins. They are also, taken collectively, a powerful source of information on scholarly influence, links between authors, and academic publishing trends. This paper will present work in progress to discover the extent to which citation patterns by UK historians have been affected by digitisation of historical newspapers. Previous studies into digital resources have used citation analysis for impact analysis (Meyer *et al.*, 2009), but faced problems gathering accurate data due to researchers' unwillingness to cite digital resources. Text mining mentions of titles within a digital resource offers a solution to this problem; indeed, Milligan (2013) has successfully used techniques from Natural Language Processing to track citations of major Canadian newspapers in Canadian PhD theses. However, there are local variations around academic practice, and cultural heritage digitisation, and to date there has been no large-scale study of digital resource citations in the United Kingdom.

This paper will present my efforts to mine newspaper citation trends using over 6,000 history theses submitted at UK Higher Education institutions, from 1999 to 2015. It will also consider the implications of text mining using legal deposit library collections. Its significance is therefore twofold. It is the first study to use text mining to track citations in UK history theses, thereby providing insights into the effect of local digitised primary sources. Second, by collaborating with British Library Labs, it provides an important test case of the possibility of exploring the text mining exception in UK copyright law. This study therefore focuses on two key research questions:

- What does citation analysis of UK history theses tell us about the impact of historic newspaper digitisation on early career historians in the United Kingdom?
- How far does the ability to text and data mine copyrighted materials provide for data driven approaches when applied to legal deposit library collections?

Research context

In the last fifteen years several studies have proposed models for evaluating the impact of digital resources (Warwick *et al.*, 2006; Meyer *et al.*, 2009; Tanner, 2012). Citation analysis is commonly used in impact evaluation, and is a well-established bibliometric technique for impact analysis (MacRoberts and MacRoberts, 1989). As part of the wider field of bibliometrics, citation analysis has been used to judge the impact of academic publications and digital resources (Meyer *et al.*, 2009). Citations are proxy measures of how frequently a document or resource is used, founded on the assumption that there is a strong positive correlation between the number of citations that a resource or article attracts, and the quality of that resource (Smith, 1980). The reality, though, is that citation practices are not always followed or understood by academics. Smith (1980) noted several reasons a scholar may choose not to cite a document: inability to obtain the document; inability to read a foreign language; lack of relevance to their work; or lack of awareness of existing work. There are further reasons for digital resources; not least a lack of awareness of how to cite them (Sukovic, 2009), and disciplinary unwillingness to acknowledge their usage (Meyer, 2009). As a result, the traditional method of mining only citations provide an unreliable picture of citation levels of digital resources.

Furthermore, there is a need to account for the varied local and national context within which researchers operate. In Canada, for instance, there is a feeling among scholars that the "limited and fragmented" (Kheraj, 2014) newspaper digitisation programme lags behind nations such as the USA, Australia and United Kingdom. The prominence of *The Toronto Star* and *The Globe and Mail* mirrors the early years of newspaper digitisation in the United Kingdom, where the ubiquity of the *Times Digital Archive* encouraged some to overstate its representativeness (Bingham, 2010). Contemporary digitised newspaper resources in the UK by contrast, tend to aggregate dozens of newspaper titles into a single resource. This paper therefore explores the likelihood that this aggregation process may have caused different citation patterns among UK researchers, who are the largest group to access these resources (Gooding, 2014), providing a comparison with the differing Canadian context presented by Milligan.

Methodology

To achieve this, I will focus on newspaper titles from two British Library resources: *British Library Nineteenth Century Newspapers (BNCN)* and *The British Newspaper Archive (BNA).* I intend to identify mentions of newspapers by title within the full text of UK history theses. The dataset comes from EThOS, a national service which makes UK doctoral theses available online for searching and reading. EThOS contains approximately 440,000 records relating to theses awarded by over 120 institutions. It provides a comprehensive, systematically collected dataset for comparison of citation trends over time. Around 160,000 records provide access to searchable full text, and I have worked with British Library Labs to identify a subset of history theses published from 1999 to 2015; in total, over 6,000 theses were identified using Dewey Decimal Classifiers, covering eight years before and after the launch of BNCN in 2007. These theses will be searched using Natural Language Processing techniques to identify the incidence of specific newspaper titles that are included in BNCN and BNA, allowing me to identify how many theses use a given source, and how frequently each source was used.

This project also acts as a test case for text mining of British Library collections. In 2014, the UK government introduced an exception to copyright law to ensure that researchers undertaking text and data-mining for non-commercial purposes would no longer infringe copyright (Intellectual Property Office, 2014), without requiring that publishers took steps to guarantee the availability of suitable datasets for text mining. In reality, this means that there are many potentially valuable data sources that could be legally studied, but no infrastructure to do so. Starting with the British Library's PhD thesis holdings, I intend to work with British Library Labs to explore the possibility of opening up further datasets for text mining.

Conclusion

This paper will explore the ways in which historical newspaper digitisation have impacted upon historiography among early career researchers in the United Kingdom, by tracking citations of digitised newspaper titles over time in the full text of over 6,000 PhD theses. This is the first study to apply text mining of digital resource citations to the UK context. It also provides an important case study of text and data mining in legal deposit library collections, in the light of current limitations to the UK copyright exception. In doing so, this project will not only illuminate the ways in which digital resources can affect local research practices, but will demonstrate the utility of text mining in addressing the methodological limitations of citation analysis for digital resources. We must adopt new computational methods to ensure that the traces of this use are not wiped away by citation practices which continue to de-emphasise the role of digital resources in contemporary research.

References

Bingham, A. (2010). 'The Digitization of Newspaper Archives: Opportunities and Challenges for Historians.' *Twentieth Century British History*, 21(2): pp. 225–231. doi: 10.1093/tcbh/hwq007

Cocteau, J. (2013). *Opium: The Diary of His Cure*. 3rd Edition. London: Peter Owen Publishers.

Intellectual Property Office (2014). *Exceptions to Copyright: Research*, *UK Government*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.

Kheraj, S. (2014). 'Canada's Historical Newspaper Digitization Problem, Part 2.' *History Matters*. http://activehistory.ca/2014/02/historical-newspaper-digitization-problem/comment-page-2/

MacRoberts, M. H. and MacRoberts, B. R. (1989). 'Problems of Citation Analysis: A Critical Review.' *Journal of the American Society for Information Science*, 40(5): pp. 342–349

Meyer, E. T. (2009). *Software Tools for Bibliometrics*, *Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*. Available at: http://microsites.oii.ox.ac.uk/tidsr/kb/49/software-tools-bibliometrics

Meyer, E. T., Eccles, K., Thelwall, M. and Madsen, C. (2009). *Usage and Impact Study of JISC-Funded Phase 1 Digitisation Projects & the Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*. Oxford: Oxford Internet Institute, University of Oxford. Available at: http://microsites.oii.ox.ac.uk/tidsr/sites/microsites.oii.ox.ac.uk.tidsr/files/TIDSR_FinalReport_20July2009.pdf

Milligan, I. (2013). 'Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010.' *The Canadian Historical Review*, 94(4): pp. 540–569

Smith, L. (1980). 'Citation Analysis.' *Library Trends*, 30: pp. 83–106

Sukovic, S. (2009). 'References to e-texts in academic publications', *Journal of Documentation*, 65(6), pp. 997–1015

Tanner, S. (2012). *Measuring the Impact of Digital Resources: The Balanced Value Impact Model*. London: King's College London. Available at: http://www.kdcs.kcl.ac.uk/fileadmin/documents/pubs/BalancedValueImpactModel_SimonTanner_October2012.pdf

Warwick, C., Terras, M., Pappa, N. and Galina, I. (2006). *The LAIRAH project: log analysis of digital resources in the arts and humanities - Final report to the Arts and Humanities Research Council*. School of Library, Archive and Information Studies, University College London. http://www.ucl.ac.uk/infostudies/claire-warwick/publications/LAIRAHreport.pdf.

Site-specific Cultural Infrastructure: Promoting Access and Conquering the Digital Divide

Andre Goodrich
andre.doodrich@nwu.ac.za
North-West University, South Africa

Gustaf Templehoff
mrgustaft@gmail.com
North-West University, South Africa

Juan Steyn
juan.steyn@nwu.ac.za
North-West University, South Africa

The ability to create and deploy networking technologies able to deliver relevant content over multiple platforms has until very recently depended on access to costly technology, infrastructure and expertise. The result has been a digital divide that is particularly acute in highly unequal economic contexts like South Africa. The potential of ICT to democratize the public realm, currently dominated by large media houses catering to individuals with the economic means to access costly data services, has not been realized.

The increased availability of cheap, open-source technologies, along with the growth of communities of users around these technologies, has made possible new kinds of digital publics beyond the abovementioned constraints.

Drawing on these technologies we have developed a freestanding platform capable of empowering users to create, publish and access their own network-dependent projects. We have done this using a solar-powered single-board computer (Raspberry Pi) and have tested its application in two projects: a site-specific digital literature project; and a virtual site-specific museum of apartheid-era forced removals. Both are based in the city of Potchefstroom in South Africa's North West Province.

Site–specific digital literature project

Site-specific digital literature, by allowing users to access place-bound, multimodal digital literature via mobile devices, opens doors for creative expression, place making as well as experiencing traditional texts in a new way. It can also contribute to a better understanding of the relationship between digital interfaces, spatiality and people and empower people in terms of what is called place-making practices (Kaye, 2000:203; Turner & Davenport, 217-220).

Byderhand (Byderhand.net, 2016), is a site-specific digital literature project launched during the 2015 Aardklop National Arts Festival in Potchefstroom, South Africa. Users accessed multimodal texts on their mobile devices through QR-codes deployed around various sites at the North-West University. *Byderhand* was awarded the Aartvark prize for a ground breaking production, and the team was approached to host similar productions elsewhere. In 2016 the *Byderhand*-project and platform were used in various educational contexts. Most recently, the *Byderhand*-team, assisted 1st additional language (Afrikaans) learners at a Potchefstroom secondary school to publish their own site-specific digital literature on the school grounds.

Although there is a demand for expansion of the *Byderhand*-project, cost and scalability are limiting factors. First, in its current form, implementing the project requires a team of dedicated members. is Second, functionality depends on the provision, maintenance and upgrading of infrastructure such as servers and physical QR-codes. On the user end, mobile data cost is a constraining factor.

Therefore, we explored the design of the infrastructure for an automated platform that could empower a low-skilled user, to create, publish and access their own projects. This was important as usability of a platform can ether promote access or restrict it through the way the interface is developed (Shneiderman, B. 2003).

The infrastructure chosen was a single-board computer (Raspberry Pi) and a platform was developed based on open standards.

Virtual museum project

South Africa's heritage landscape remains largely skewed in ways that exclude intangible heritage such as stories and memories of such events as apartheid-era forced removals. A virtual museum of forced removals aids in remedying this by offering an in situ space where people can give voice to their experience and preserve it for future generations.

The city of Potchefstroom was one of the first places where forced removals took place during the apartheid regime but little historical records exist outside oral accounts and some pictures of what took place. The platform we developed provides access for a community to tell their own story without the need for expensive infrastructure. It can provide a place for reconciliation, healing and understanding the intricacies of racial tensions within the context of a specific community.

In summary our paper reports on prototyping a cheap, scalable system, independent of electricity and data costs that can allow users to add and access content for both these projects. In particular, the paper considers the following:

1. How the prototype platform enables users to publish multimodal texts and multimedia (e.g. audio, images and video) on their own.
2. How the prototype platform provides an easy way to consume published work using entry-level smartphones.
3. The scalability of this cultural infrastructure, with regard to expanding multimodal site-specific literature and museums along with the creation of site-specific corpora.
4. Educational applications:
    a. How the platform could be ported into other applications such as freestanding, free-access libraries for schools and community centers.
    b. Educational possibilities created that can introduce students to the world of digital literature.
    c. A lesson in the value of interdisciplinary collaborations.
5. Building collaborative communities and empowering them to tell their own stories.
6. Mini-showcase of the project

Bibliography

byderhand.net (2016). Interaktiewe Leeservaring http://byderhand.net/

Gong, J. and Tarasewich, P. (2004) Guidelines for handheld mobile device interface design. In *Proceedings of DSI 2004 Annual Meeting* (pp. 3751-3756).

Kaye, N. (2000) *Site-Specific Art: Performance, Place and Documentation*

Shneiderman, B. (2003) "*Promoting universal usability with multi-layer interface design.*" In ACM SIGCAPH Computers and the Physically Handicapped (No. 73-74, pp. 1-8). ACM.

Turner, P. & Davenport, E. *eds.* (2006) *Spaces, Spatiality and Technology*

Designing Tools for Macro-Scale Data Analysis in the History of Science

Elyse Graham
jean.graham@stonybrook.edu
SUNY Stony Brook, United States of America

Robert Crease
robert.crease@stonybrook.edu
SUNY Stony Brook, United States of America

Our project involves developing a new kind of digital resource to capture the history of research at scientific facilities in the era of the "New Big Science." The extent, scope, and diversity of research at such facilities makes keeping track of it difficult to compile using traditional historical methods and linear narratives; there are too many overlapping and bifurcating threads. In this talk, we will discuss existing methods of data collection and curation for a specific case project, the National Synchrotron Light Source Digital Archive. We are especially interested in the functional potential, in the context of this kind of tool development, of the humanistic concepts of narrative, metaphor, and performance.

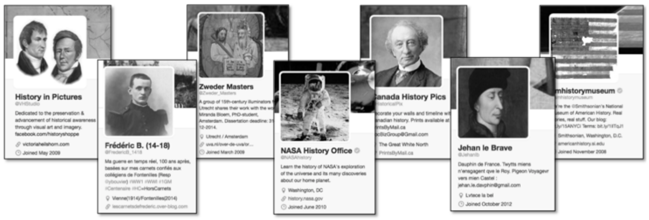
Médias sociaux et mise en scène de l'histoire

Martin Grandjean
martin.grandjean@unil.ch
Université de Lausanne, Switzerland

Introduction

La rigueur de la critique des sources historiques est-elle soluble dans l'immédiateté et la course à l'audience des médias sociaux ? Depuis plusieurs années, les projets de médiation culturelle se multiplient, en particulier sur Twitter où la brièveté des messages se prête bien à une nouvelle forme de *storytelling* historique et dont le public est particulièrement curieux et avide de *retweeter* l'une ou l'autre photographie d'archive documentant la grande (et plus rarement la petite) histoire de l'humanité ou de participer à la commémoration d'un événement (Clavert *et al.* 2015). Mais les institutions patrimoniales et les universités ne sont pas les seules à mettre ainsi en valeur leurs collections et leurs compétences, elles sont aujourd'hui largement dépassées par des comptes créés de toutes pièces pour partager automatiquement des images célèbres ou des citations flatteuses à des millions d'abonnés, sans contextualisation ni précautions de véracité, de licence ou de liens vers la source originale. Alors que certaines de ces initiatives sont le fait de passionnés, qui s'attachent à un événement qui les touchent personnellement et ont à coeur de le faire vivre à leurs abonnés, nombre de pages Facebooks/Twitter sont en fait produites en série par des individus qui, conjointement à des comptes partageant les plus belles photos de paysages du monde, d'animaux ou de personnalités célèbres, construisent des audiences de dizaines de millions d'internautes à des fins publicitaires.

Nous proposons une analyse d'un certain nombre de comptes Twitter présentant des caractéristiques très différentes. On verra que la montée en puissance des comptes commerciaux est basée sur des stratégies relativement simples, de l'achat de *followers* à l'auto-référencement. Dans un second temps, nous établirons une typologie des usages des documents historiques à des fins de communication sur les médias sociaux. Ce classement, qui se veut également un outil pour les institutions qui seraient tentées de mettre au point leur propre projet de médiation culturelle sur ces plateformes, montrera en particulier que la palette des usages ne se restreint pas forcément au partage de ressources mais tient parfois beaucoup plus du récit et de la mise en scène.



History in Pictures
@VHStudio
Dedicated to the preservation & advancement of historical awareness through visual art and imagery. facebook.com/historyshoppe
victoriahelshorn.com
Joined May 2009

Frédéric B. (14-18)
@FredericB_1418
Ma guerre en temps réel, 100 ans après, basées sur mes carnets confiés aux collégiens de Fontenilles (Resp @ybouvier) #WW1 #WWI #1GM #Centenaire #HC#HorsCarnets
Vienne(1914)/Fontenilles(2014)
lescarnetsdefrederic.over-blog.com

Zweder Masters
@Zweder_Masters
A group of 15th-century illuminators f
Utrecht shares their work with the wo
Miranda Bloem, PhD-student,
Amsterdam. Dissertation deadline: 31
12-2014.
Utrecht / Amsterdam
uva.nl/over-de-uva/or...
Joined March 2009

NASA History Office
@NASAhistory
Learn the history of NASA's exploration of the universe and its many discoveries about our home planet.
Washington, DC
history.nasa.gov
Joined June 2010

anada History Pics
istoricalPix
ecorate your walls and timeline wi
anadian history. Prints available at
intsByMail.ca
cBizGroup@Gmail.com
The Great White North
PrintsByMail.ca

Jehan le Brave
@Jehanlb
Dauphin de France. Twytts miens n'engagent qve le Roy. Pigeon Voyagevr vers mien Castel : jehan.le.davphin@gmail.com
Lvtece la bel
Joined October 2012

mhistorymuseum
mhistorymuseum
'e the @Smithsonian's National
seum of American History. Real
ries, real stuff. Our blog:
ly/15ANYCI Terms: bit.ly/1lTqJ1
Smithsonian, Washington, D.C.
americanhistory.si.edu
Joined November 2008

Figure 1. Exemples de comptes Twitter à caractère historique.

Prendre la mesure d'un phénomène

Un compte Twitter comme @HistoryInPics génère quotidiennement plus de 10,000 retweets, alors que les *social media editors* de la Bibliothèque nationale de France (@GallicaBnF) dépassent occasionnellement la centaine d'interactions, malgré leurs contenus originaux et souvent présentés avec humour. Mais plus d'abonnés ne veut pas toujours dire plus d'engagement de ceux-ci, surtout quand une majorité d'entre eux sont inactifs : c'est pourquoi nous analyserons en détail les publications d'une dizaine de comptes pendant un mois. La quantification de ces audiences (fig. 2) est un passage obligé pour comprendre l'ampleur du phénomène et tenter de comprendre la nature de ces publics. Fondamentalement, si ces démarches – même commerciales – ont un tel succès, c'est qu'il est temps que les historiens se saisissent de la question.

On verra en particulier que les audiences des grands comptes peu scrupuleux se recoupent largement, un phénomène qu'on expliquera en détail par l'analyse d'un échantillon de plusieurs dizaines de milliers d'abonnés. On s'attachera également à dresser, pour les comptes d'institutions patrimoniales, une statistique des contenus les plus susceptibles de créer un fort engagement. Encore une fois, il s'agit d'adopter une posture d'observation, de décoder des pratiques numériques souvent en décalage avec les idées reçues qui circulent dans les milieux académiques et de questionner la relation ambigüe entre recherche historique et valorisation de celle-ci.

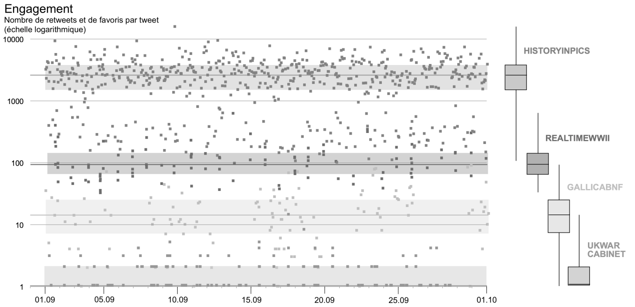


Figure 2 : Analyse d'un mois de tweets de quatre comptes Twitter partageant des contenus historiques de manière très différente (Grandjean 2014). Chaque point est un tweet dont le succès est représenté sur une échelle logarithmique de retweets+favoris. Les boîtes de droite résument la dispersion du résultat (50% des tweets de chaque compte obtient un "engagement" qui tient dans la zone définie par la boîte correspondante).

Décrire les usages

Certains mettent valeur leurs collections, d'autres en font des exercices (Steffen et Nunes Coelho 2014), cherchent le *buzz* pour lancer une start-up en communication numérique, luttent pour qu'une cause historique soit reconnue ou mettent en scène leur grand-père au travers de son journal de guerre, mais tous ont en commun l'utilisation de documents d'archives – souvent photographiques – dans leurs micro-messages.

Dans cette recherche, nous ferons la différence entre les communications basées sur la disponibilité d'archives et celles qui, pour illustrer un récit, font usage de documents d'archive (fig. 3). Paradoxalement, c'est dans la première catégorie que l'on trouve les usages les plus « purs » comme les usages les plus critiquables : d'une part des institutions d'archive qui partagent des pièces de leurs collections ou des chercheurs qui consultent des documents et tweetent les « perles » qu'ils rencontrent et d'autre part des comptes semi-automatisés qui partagent des photographies issues de grandes banques d'images, du portrait de président américain à la couverture d'album des Beatles. Il s'agit toujours ici d'un processus archive→communication. À l'inverse, la démarche communication→archive est mise en pratique lorsque l'internaute souhaite mettre en scène un événement, un récit historique ou une thématique et que celui-ci va piocher dans des illustrations ou des documents pour étayer son propos.

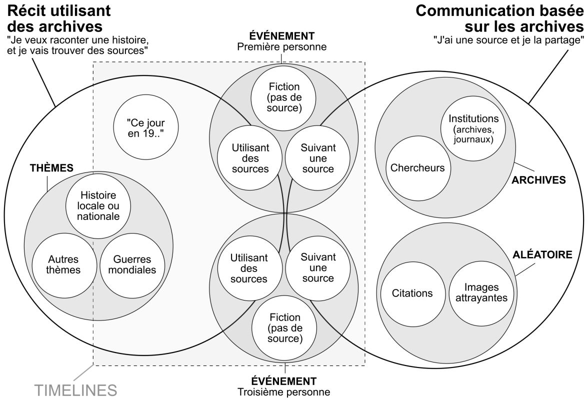


Figure 3 : Typologie des usages de Twitter pour la communication de documents historiques.

Perspectives

Les médias sociaux favorisent-ils une démocratisation et une réappropriation de l'histoire par son public ou sont-ils au contraire un vecteur d'une histoire spectacle décontextualisée ? S'il apparaît de manière évidente que la réponse se situe entre ces deux extrêmes, ou du moins qu'elle varie selon les usages précis, rappelons que nous proposons ici une plongée dans une réalité, quotidienne pour certain et étrangère pour d'autres. Cet état des lieux, en deux parties, quantitative et typologique, doit servir de base à une réflexion plus large sur le rôle des historiens et des institutions patrimoniales dans une société numérique où l'on ne peut lutter contre la dissémination d'images mal référencées et de contenus instrumentalisés. Cerner ces phénomènes, c'est aussi préparer la réplique, rigoureuse, critique et pourquoi pas créative.

Bibliography

Butticaz E. (2013). Twitter, cette machine à remonter le temps, *Le Temps*, https://www.letemps.ch/no-section/2013/11/22/twitter-cette-machine-remonter-temps (accessed 1st October 2016).

Clavert F., Majerus B. et Beaupré N. (2015). Twitter, the Centenary of the First World War and the Historian. *Twitter for Research 2015*, Lyon.

Grandjean M. (2014). Source criticism in 140 characters: rewriting history on social networks, *International Federation for Public History Conference*, Amsterdam.

Steffen M. et **Nunes Coelho P**. (2014). Tweeting during World War II, http://h-europe.uni.lu/?p=2037 (accessed 1st October 2016).

Varin V. (2014). Tweeps Discover the Past, *Perspectives on History*, https://www.historians.org/publications-and-directories/perspectives-on-history/april-2014/tweeps-discover-the-past (accessed 1st October 2016).

New potentials in the digital archives: a participative inquiry into interdisciplinary collaboration in digital historical research at the Wellcome Trust

Alex Green
a.green@wellcome.ac.uk
Wellcome Trust, United Kingdom

Lalita Kapish
l.kaplish@wellcome.ac.uk
Wellcome Trust, United Kingdom

Hannah Walker
h.walker@wellcome.ac.uk
Wellcome Trust, United Kingdom

This paper discusses an exploratory project where a group of university academics, software developers, designers and librarians spent a week analysing a broad selection of the Wellcome Library's digital collections with an aim to explore new ways of conducting collaborative digital history research, identifying and documenting barriers and successes and also pointing to gaps in institutional support infrastructures.

Over the past two decades, significant quantities of cultural heritage have been digitised. Internet Archive has digitised over 20 million items in partnership with libraries and collections around the world. Google has digitised over 25 million books as Google Books. Alongside this the emergent field of digital humanities has sought to take advantage of new opportunities afforded by unprecedented access to collections. Numerous commentators and researchers have, in the words of Hayles (2015), voiced the opinion that "if there is an area of general agreement, it is the transformative potential of digital humanities for the humanities and for academic discourse" (see also Ogilvie, 2016; Alves, 2014).

Wellcome Library, part of the Wellcome Trust, is one of the world's most significant collections relating to health and medicine, with works ranging from posters and paintings to personal archives, printed books and packaging ephemera. Through its digitisation programmes, Wellcome is a major producer of digitised historical material and datasets, with over 40,000 digitised archives, nearly 100,000 digitised monographs and over 10,000 artworks, manuscripts, videos and reports. These have been made freely available under the most liberal licence possible, dependent on the copyright status of the material. In addition, IIIF image services (IIIF Consortium) including standardised image and presentation APIs, along with services for OCR and text search are provided (Chaplin, 2016).

A key aim for Wellcome is to enable new types of research and knowledge production; our mission is 'to improve health for everyone by helping great ideas to thrive' (Wellcome Trust, 2016). One way we seek to do this is through enabling the exploration of cultural and social meanings of health in the past. Digitisation has undoubtedly increased use of our collections significantly and developed a large international audience, with over 50% of researchers accessing content from outside the UK. To better understand the users and usage of our digital collections we have carried out quantitative and qualitative research in collaboration with Prof. Pauline Leonard (Green and Andersen 2016). Our findings showed that beyond increasing access, the nature of collection usage has not changed significantly. The majority of researchers still access works within a single collection, page through digitised works in a linear fashion, make limited use of OCR search and little to no use of APIs. We are not yet seeing Hayle's "transformative potential" (2015) realised. To better understand this, we developed a number of hypotheses and research questions, which we categorise here under four loose umbrellas:

Lean working

In summarising the findings of multiple digital historical studies, Alves has argued that they 'implicitly confirm the efficiency of digital means… but also… that their application is, often, generally associated with expensive projects requiring extensive human resources with diverse skills' (2014). We questioned whether digital research is necessarily expensive or requiring of extensive resource. Drawing inspiration from commercial software development, we asked what is the Minimum Viable Product (Leanstack, 2016); can teams achieve meaningful research relatively inexpensively through an agile approach of iterative investigation and identification of emergent areas of interest rather than pre-defining fixed research questions.

Skills and knowledge

It could be argued that there is a skills or knowledge gap within traditional historical research communities which inhibits conducting digital research. However, we questioned if this was truly a barrier when evidence from citation patterns shows increases in collaborative approaches to digital humanities research (Nyhan and Williams, 2013). We questioned if extending the scope of traditional research teams to include commercial development partners could bring new insights and capabilities.

Crossing collections

Wellcome's digital collections include intensely heterogeneous material along with data sourced from multiple institutions. We questioned whether there are practical or technical barriers to break down divisions between collections and drawing from a range of content. Hitchcock (2013) has argued that 'the lack of flexibility of the available digital tools [has] enabled only the effective utilization and analysis of quantitative sources or sources easily transformed into a quantitative format'. As many of Wellcome's collections are archival and contain large quantities of handwritten and pictorial material, we specifically wanted to explore the possibilities of digital for non-quantitative research, and research which still requires 'close reading' of sources (Van Dijk, 1985).

Quality and consistency

Areas of enquiry in this umbrella included whether our digital collection is suitable in its scale and scope, and the quality and consistency of OCR and collection metadata. We also questioned if we we had the right kind of web services available and their usability for individuals with different levels of experience.

To explore and better understand these questions we designed an experimental approach, adapting a concept becoming familiar to cultural institutions – the 'hack day' – and extending it out into a week-long intensive R&D project, where small teams led by a mix of independent and academic researchers would work collaboratively to explore the Wellcome Library digital collections. Our participants included research staff from several UK universities, librarians and archivists, commercial designers and software developers. Through careful screening of participants, we selected researchers with shared enthusiasm for, but variable experience of digital historical research. This choice was deliberate in order to focus in on barriers relating to experience, skills and technical feasibility without confounding these variables with any reluctance to use digital methods. However, the enthusiasm for digital methods in the historical community is an important question and undoubtedly merits further investigation. Research areas were open-ended, with a focus on experimentation rather than production of finished work. However, we did agree broad areas, including handwritten records from a private asylum, 5000 Medical Officer of Health reports covering London from 1848-1972, 6,600 issues of the trade journal Chemist and Druggist and the 79,000 books digitised by the UK-MHL project (Wellcome Library, 2016).

Drawing from approaches of participative and cooperative inquiry (Reason and Bradbury, 2008) we embedded Wellcome staff in mixed teams of academic staff, developers and designers, positioning the teams as researchers of the collections, participants in a broader experiment of research production and observers and documenters of the experiment itself. We encouraged self-documentation by teams through use of project boards, blogs and wikis, conducted individual interviews with participants throughout the process and held daily reviews and a plenary session to discuss progress and reflect on the experiment.

During the week we produced multiple tools and visualisations, and also historical findings. These are documented on the project blog, along with the processes each team went through. This clearly demonstrated the potential of the collections, but also the barriers to working with them. Interestingly, the project using digitised archival material without OCR was one of the most successful – partly due to the synergy of the team members, but also the clarity of the challenge for the material. Two of the five projects seeded in the week continue to be developed, and we are continuing to provide support to the researchers leading them.

Feedback from participants in the project identified particularly the benefits of working in teams with mixed skill-sets. Developers and library staff gained from acquiring greater understanding of research processes and interests, while researchers gained access to technical skills, and exposure to a different approach to digital working. One participant remarked: 'I found working with other people, from a variety of different backgrounds, really generative – both for thinking about why people who do different kinds of work approach digital resources in different ways, and for thinking about my own research along new lines.'

The web services we provide for programmatic access to digital collections were identified as a particular barrier during the week. While the images and bibliographic metadata for our collections are exposed through an international standard, the Image Interoperability Framework (IIIF), this proved conceptually complex for developers to quickly prototype with, requiring combining and processing multiple JSON responses to access digital items. Adding further complexity to this, our collection OCR is available primarily through ALTO XML (Library of Congress, 2016) which our developers found challenging to process at scale.

As cultural heritage institution and a research funder, we are continuing to unpack the implications of the findings. As a library, we found great value in co-production with research users so will be repeating a similar annual event, investigating different aspects of our collections. We have also identified particular issues related to the usability of our collections and added these to our development roadmap. As a funder we are considering options for increasing innovation by seeding early stage research through similar collaborative processes. To take this forward we will be running a series of pilot events through our interdisciplinary research residency, The Hub.

Bibliography

Alves, D. (2014). Guest Editor's Introduction: Digital Methods and Tools for Historical Research. *International Journal of Humanities and Arts Computing*. **8**(1):1-12.

Bergold, J. and Thomas, S. (2012). Participatory Research Methods: A Methodological Approach in Motion. F*orum Qualitative Sozialforschung / Forum: Qualitative Social Research*. **13** (1): Art. 30, http://nbn-resolving.de/urn:nbn:de:0114-fqs1201302.

Chaplin, S. (2015). Why Creating a Digital Library for the History of Medicine is Harder than You'd Think! *Medical History*. **60**(01):126-129.

Green, A & Andersen, A. (2016). 'Finding value beyond the dashboard', paper presented to MW2016: Museums and the Web 2016, Los Angeles, 6-9 April. Available at http://mw2016.museumsandtheweb.com/proposal/guaging-value-beyond-the-dashboard/ (Accessed: 1 Novemeber 2016).

Hayles, N. (2015) How We Think: Transforming Power and Digital Technologies. In Svensson, P. and Goldberg, D. (eds), *Between Humanities and the Digital*. MIT; 2015. p. 503.

Hitchcock, T. (2013). Confronting the Digital. *Cultural and Social History*. **10**(1):12-14.

IIIF Consortium. (no date). *International image Interoperability framework*. Available at: http://iiif.io/about/ (Accessed: 1 November 2016).

Library of Congress. (2016). *About ALTO.* Available at: http://www.loc.gov/standards/alto/about.html (Accessed: 5 April 2017)

Leanstack. (2016). *Minimum viable product*. Available at: https://leanstack.com/minimum-viable-product/ (Accessed: 1 November 2016).

Ogilvie, B. (2016). Scientific Archives in the Age of Digitization. *Isis*. **107**(1): 77-85.

Nyhan, J. and Duke-Williams, O. (2016). Joint and multi-authored publication patterns in the Digital Humanities. *Arche Logos*. http://archelogos.hypotheses.org/103

Reason, P. & Bradbury, H. (2008). Introduction. In Reason, P. and Bradbury, H.(eds), *The Sage handbook of action research. Participative inquiry and practice.* London: Sage, pp.1-10.

Toon, E, Timmermann, C & Worboys, M. (2016). Text-Mining and the History of Medicine: Big Data, Big Questions?. *Medical History*. **60**(02):294-296

van Dijk, T.A. (ed.) (1985). *Handbook of discourse analysis: V. 1: Disciplines of discourse*. 3rd ed. Orlando: Academic Press.

Wellcome Library. (2016). *UK medical heritage library*. Available at: https://wellcomelibrary.org/collections/digital-collections/uk-medical-heritage-library/ (Accessed: 1 November 2016).

Wellcome Trust. (2016). *Our Strategy*. Available at: https://wellcome.ac.uk/about-us/our-strategy (Accessed: 1 November 2016).

Dazzle them with Baffles: Gauging Attitudes toward Digital Fabrication in an Online Musicians' Community

Brian Greenspan
brian.greenspan@carleton.ca
Carleton University, Canada

This paper presents a project, currently underway, that combines 3D fabrication and digital audio processing with text mining and sentiment analysis to gauge attitudes toward digital fabrication processes (and maker culture more generally) among a large online community of musicians.

In the last two years, traditional musical instrument manufacturing materials have been joined by laser-sintered plastics and photopolymer resins, as new Additive Manufacturing techniques have been employed either to replicate vintage instrument parts, or redesign them altogether. In no case, however, have digital fabrication methods managed to entirely displace artisanal tradition.

This study explores how digital fabrication has been negotiated by one large and active online musician's community that values artistry, craft, and industrial history alike. Using low-end desktop 3D fabrication methods with open-source software, I am reproducing a line of replica instrument parts based on vintage originals. Once all reproductions are complete, I plan to engage a professional musician to test-play the printed replicas, and compare them with their original vintage models. These tests will be recorded using audio freeware, and shared with members of a select online musician's community, in order to determine whether they distinguish my 3D-prints from their original models.

My aim is less to determine whether these replicas are in fact morphologically and sonically identical to their period models than to determine how 3D digital fabrication is received by musicians strongly devoted to vintage equipment and artisanal craft. Using the Python Natural Language Toolkit (NLTK 3.0) with a naïve Bayes classifier trained on a valenced wordlist (e.g. AFINN), I will conduct sentiment analyses of the online forum to gauge how the community of musicians responds to the introduction of new manufacturing techniques that are neither industrial nor conventionally artisanal.

Ultimately, this project addresses the "emergence" of digital humanities (Jones 2013) into artisanal practices as a sign of our contemporary "post-digital" condition, under which we need "no longer talk about digital versus analog but instead about modulations of the digital or different intensities of the computational" (Berry 2014). By engaging one active and prominent community of analog musicians in a discussion of digital fabrication, I hope to use their insights into the materiality of music production to address questions crucial to maker epistemology, such as: "For which methodologies do tactility and texture especially matter?," and "When is scholarly communication most persuasive off the screen?" (Sayers 2015).

Bibliography

Berry, D.M. (2014). "Post-Digital Humanities: Computation and Cultural Critique in the Arts and Humanities." Educause Review 49.3: 22-6.

Jones, S.E. (2013). The Emergence of the Digital Humanities. New York: Routledge.

Sayers, J. (2015). "Why Fabricate?" Scholarly and Research Communication 6.3: 1-11.

Linked Places: A Modeling Pattern and Software for Representing Historical Movement

Karl Grossner
karlg@worldheritageweb.org
World Heritage Web, United States of America

Merrick Lex Berman
mberman@cga.harvard.edu
Harvard University, United States of America

Rainer Simon
rainer.simon@ait.ac.at
Austrian Institute of Technology, Austria

Introduction

This paper reports on work in progress aimed at facilitating the creation, sharing, linking, and analysis of data about the movement of people, ideas, cultural practices, and commodities between places, over the course of history. Products of the Linked Places project include: conceptual and logical models for historical routes; a temporal extension of the popular GeoJSON data format, called GeoJSON-T; several varied exemplar data sets converted to GeoJSON-T format; prototype web software for browsing and visualizing that data; and Python scripts to convert data between CSV, GeoJSON-T, and RDF compatible with the Pelagios Gazetteer Interconnection Format. Substantial interim work products are shared in the Linked Places and GeoJSON-T GitHub repositories and have been reported in some detail in two blog posts (1, 2) .

Motivation

A growing number of historical gazetteers are being developed in the course of digital humanities research projects (Berman, Mostern & Southall, 2016). Their spatial temporal coverage is typically limited to a particular area and period due to factors of scholarly quality, cost, and relevance to a given project. Coverage extents do vary considerably, from a single city for a few generations to a region for several centuries. With few exceptions, these gazetteers are unpublished as such; instead they are spatial tables contained within, and integral to, the larger project data store.

Because historical gazetteers are difficult and time-consuming to produce, it is vital they be published, when possible, in a way that permits linking them—an activity that the Pelagios project has made great strides in facilitating. An emergent network of specialized gazetteers holds terrific promise, not only for re-use, but ultimately as a distributed, increasingly comprehensive geographical (i.e. spatial-temporal) index to linked data from numerous domains, including history, archaeology, literary studies, philology, and several of the social sciences. The focus of such an index, and encyclopedic applications it enables, will be on individual places, typically at the scales of cities and points of interest.

Such systems are highly desirable, but given a large volume of data about individual places we can also begin harvesting, creating, and sharing data about the connections between them. We should be able to ask of historical gazetteers: What journeys and historical routes has a given place been a waypoint on? And, what flows of people, ideas, and commodities has it been a source or sink for?

But the Linked Places and GeoJSON-T projects have been undertaken with an even larger, “moonshot” vision in mind: a system allowing scholars and the general public to visualize and analyze the emergence, growth and spread of human settlements, their changing attributes, and the dynamic connections between them, including the diffusion of technologies and cultural practices.

To realize these ideas, we need a) lots of data, and b) methods and means for merging or linking them. In some respects, we are starting from scratch; data about historical movement is sparse and stored in disparate forms. Much of it will be newly generated, for example by parsing texts, transforming tabular records, or digitally tracing lines on historical maps. Merging and linking operations will require that the form of data from different sources (or abbreviated catalogues thereof) be either standardized (in the case of merging), or similar enough that automated alignment is feasible.

The majority of works on geographic networks concerns physical media like roads and rail, whereas movement data is eventive. Geographers have modeled migration flows and disease diffusion for several decades, providing theoretical bases for their analysis that are outside our present scope. An overview of that work is found in (Lowe & Moryadas 1975). An excellent and more recent work on mobility and geographic movement is Tim Cresswell’s “On the Move” (2006). We are not aware of any efforts to model data for historical routes computationally, however the core abstraction we build upon is the traditional graph/network model of nodes and edges credited to 18th century work of Euler (Biggs, et al 1986).

A Modeling Pattern

Data modeling is as much an art as a science (Simsion & Witt 2004), but some core best practices are well-known. A typical first step is establishing what entities are to be represented, what their essential attributes are, and what relationships obtain between them (cf. Chen, 1976). This step is often best accomplished collaboratively, in an iterative process undertaken by domain experts. Our results were immediately published to blog posts and relevant listservs, and the resulting input was useful in refining the model.

When the modeling context is an individual research project, it hardly matters what names are given those entities and relationships—only that the data store's internal logic be sound and well understood by project members. But if, as in this case, the system will accommodate data from many sources or be accessed by others, we need to find broad agreement on a conceptual model and a vocabulary for its constituents between as many prospective participants as possible—that is, to describe the ontology of the research domain. Although much ontology engineering of this sort has involved comprehensive high-level ontologies such as the CIDOC-CRM , the development and implementation of small ontology design patterns (ODP) has been gaining favor since the introduction of that paradigm by Aldo Gangemi (2005). Such patterns, by any name, are "reusable successful solutions to a recurrent modeling problem" (definition provided by the Association for Ontology Design & Patterns (ODPA)) which can be used alone or assembled in modular fashion for larger requirements. Examples include patterns for "Place," "Event," "Participation," and "Region."

And so the first step taken in the Linked Places project has been to develop an ontology design pattern for the historical movement of something between two or more places over some physical channel, either for some time during or throughout a timespan. The pattern, visualized in Figure 1, comprises the following conceptual understandings:

A route describes an attestation of one or more occurrences of the movement of something (e.g. people, commodities, information) between two or more places, either for some time during or throughout a time_period. Routes are composed of one or more segment, each of which is composed of two places and a path (corresponding to nodes and edges in network parlance), the locations and temporal attributes for which may be unknown or unspecified. Movement between places occurred upon ways (the term used by OpenStreetMap) —physical channels such as roads, rivers, canals, railways, footpaths, and sea lanes—and may have been directional.

The three types of routes considered here are journeys, flows, and historical_routes:

A journey is the record of a specific instance of travel by one or more individuals. Examples include: the 7th century pilgrimage of the Buddhist monk Xuanzang across China and India; the first voyage of Captain James Cook, between 1768 and 1771.

A flow is the record of the movement of something (commodities, people, ideas) between two places, aggregated as a magnitude over a period of time. Examples include: the transport of captive Africans between West Africa and Bahia in the 17th century; letters between certain correspondents in Paris and Prague in the 18th century; a source network of late Neolithic obsidian artifacts and known source locations on the Anatolian Plateau.

A historical_route asserts a single or composite named course of travel between places, taken repeatedly by unspecified individuals over time, usually for purposes of commerce. Examples include the Silk Road and the Amber Routes. Some correspond with named roads, for instance the Via Salaria in Italy is both a way and a historical_route. Additional axioms indicated by the relations and cardinality expressions (e.g. 0…*) in Figure 1 include:

- All routes are sourced, normally to textual or cartographic documents
- The way for a segment (its physical path described by a geometry) may be known and represented, unknown, or ignored (Segments with unspecified ways will typically be visualized as a line or arc)
- Each segment has one or more temporal attribute ("when"), which can be a time_period, (possibly named) or a sequence (e.g. after segment n)
- Routes and their component segments can have any number of attributes (properties), dependent upon data sources and project requirements

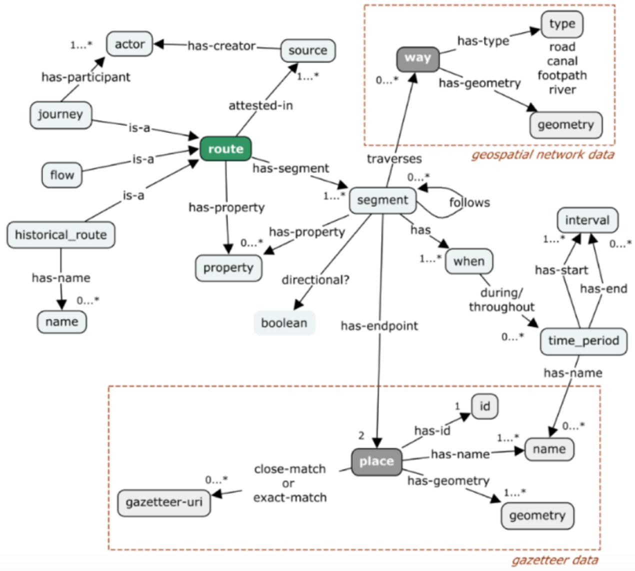
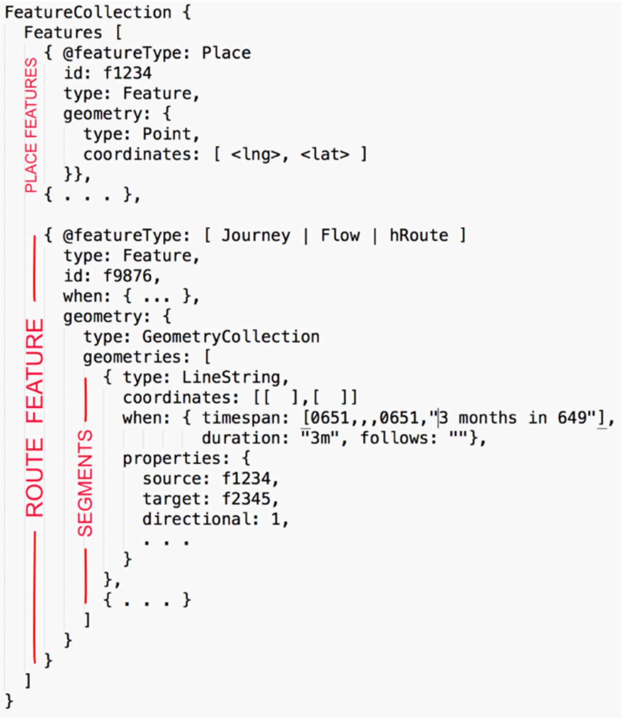


Figure 1. A conceptual model for historical movement (routes)

The ontology pattern we introduce here is specialized, as compared to high level ontologies like CIDOC-CRM. We have not yet mapped our distinctive entities (route, journey, flow, historical_route, segment, when) to existing ontologies. The term place is commonly found, but usually is synonymous with location; the sense we are adopting is that of the Pleiades gazetteer, but is not in a published ontology that we're aware of. In any case, we feel it is best to first lay down a logically coherent set of terms and at a later date attempt to align them with other ontologies.

Formats

The route ODP has informed our development and implementation of recommended standard data formats. It turns out all three types of routes can be effectively described in GeoJSON-T, an extended version of GeoJSON, the widely-used format for representing geographic FeatureCollections. A FeatureCollection of routes will include both Place and Route features. Route segments are articulated as an array of one or more geometries in a route's GeometryCollection. GeoJSON-T allows optional "when" objects, both for each feature at the same level as its geometry object and for segment geometries (Figure 2). Features and segments have certain required properties as shown, and can have unlimited project-specific properties.



```
FeatureCollection {
  Features [
    { @featureType: Place
      id: f1234
      type: Feature,
      geometry: {
        type: Point,
        coordinates: [ <lng>, <lat> ]
      }},
    { . . . },

    { @featureType: [ Journey | Flow | hRoute ]
      type: Feature,
      id: f9876,
      when: { ... },
      geometry: {
        type: GeometryCollection
        geometries: [
          { type: LineString,
            coordinates: [[  ],[  ]]
            when: { timespan: [0651,,,0651,"3 months in 649"],
                    duration: "3m", follows: ""},
            properties: {
              source: f1234,
              target: f2345,
              directional: 1,
              . . .
            }
          },
          { . . . }
        ]
      }
    }
  ]
}
```

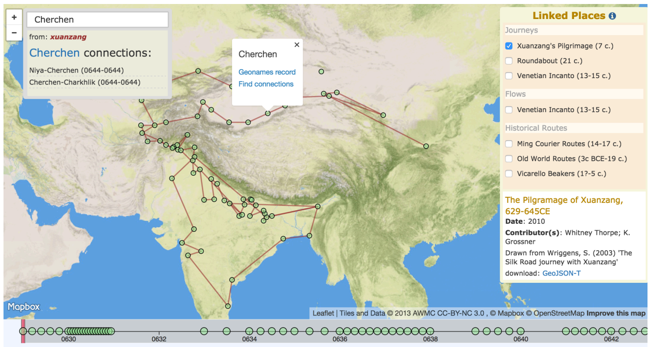
Figure 2. GeoJSON-T applied to route data

Data

To date, seven exemplar datasets have been converted from a typical CSV format to GeoJSON-T, using a newly developed Python program. Three are for journeys: two by individuals (a 7th century pilgrimage and a modern circumnavigation), the third by 840 Venetian ship convoys in the 13-15th centuries. Another dataset aggregates those ship journeys as flows having magnitudes of journeys and ships. The last three are historical_routes: the Roman era itinerary of the Vicarello Beakers, the route system between courier stations in Ming Dynasty China, and a large set of "Old World" trade and pilgrimage routes .

Software

The widespread adoption of GeoJSON has demonstrated that for a data format to be useful, there must be software with visualization and analysis capabilities that supports it. Accordingly, an essential element of the Linked Places project is development of proof of concept web software to render GeoJSON-T data, for both routes and places alone, to a map and timeline together. The development of that software is ongoing, and publicly available. (Figure 3).



+
−
Cherchen
from: xuanzang
Cherchen connections:
Niya-Cherchen (0644-0644)
Cherchen-Charkhlik (0644-0644)
Cherchen
×
Geonames record
Find connections
Linked Places
Journeys
☑ Xuanzang's Pilgrimage (7 c.)
□ Roundabout (21 c.)
□ Venetian Incanto (13-15 c.)
Flows
□ Venetian Incanto (13-15 c.)
Historical Routes
□ Ming Courier Routes (14-17 c.)
□ Old World Routes (3c BCE-19 c.)
□ Vicarello Beakers (1?-5 c.)
The Pilgramage of Xuanzang, 629-645CE
Date: 2010
Contributor(s): Whitney Thorpe; K. Grossner
Drawn from Wriggens, S. (2003) 'The Silk Road journey with Xuanzang'
download: GeoJSON-T
Mapbox
Leaflet | Tiles and Data © 2013 AWMC CC-BY-NC 3.0 , © Mapbox © OpenStreetMap Improve this map
0630
0632
0634
0636
0638
0640
0642

Figure 3. Linked Places interface (partial view as of March 2017)

Bibliography

Berman, M.L., Mostern, R., & Southall, H. (2016). Placing Names: Enriching and Integrating Gazetteers. Bloomington: Indiana University Press.

Biggs, N.; Lloyd, E.; Wilson, R. (1986), Graph Theory, 1736-1936, Oxford: Oxford University Press

Chen, P. P. S. (1976). The entity-relationship model—toward a unified view of data. ACM Transactions on Database Systems (TODS), 1(1), 9-36.

Cresswell, T. (2006). On the move: Mobility in the modern western world. New York: Routledge.

Gangemi, A. (2005). Ontology design patterns for semantic web content. In International semantic web conference (pp. 262-276). Springer Berlin Heidelberg.

Lowe, J. C., & Moryadas, S. (1975). The geography of movement. Boston: Houghton Mifflin

Simsion, G., & Witt, G. (2004). Data modeling essentials. San Francisco: Morgan Kaufmann.

Los Hilos De Ariadna En El Laberinto Temático: Visualización Y Minado De Datos Para Bibliotecas

Silvia Eunice Gutiérrez De la Torre
segutierrez@colmex.mx
El Colegio de México A.C., México

Julián Alberto Equihua Benítez
julian.equihua@gmail.com
CONABIO, México

Micaela Chávez Villa
mch@colmex.mx
El Colegio de México A.C., México

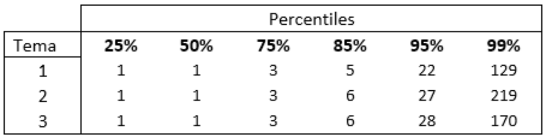
Introducción

Encontrar relaciones entre los encabezamientos que se asignan a una obra monográfica es un problema histórico en el ámbito de búsqueda y recuperación de información. Por un lado, los documentos rara vez pueden ser representados con un solo tema; por otro, el número de temas que se puede asignar a una obra es virtualmente infinito (Green, 2001). En la intersección de las Humanidades Digitales y la Bibliotecología han existido diversos esfuerzos por mejorar la calidad de las ontologías de estos temas (Nurmikko-Fuller et al, 2016), su evaluación (Harper, 2016) y visualización (Duguid, 2015). Sin embargo, a nuestro conocimiento, no se han hecho estudios que aprovechen métodos innovadores para indagar relaciones entre los encabezamientos de materia. En esta comunicación breve, presentamos los resultados preliminares de un primer acercamiento al tema, que aprovecha el área de especialidad de cada participante del equipo --humanidades digitales, ciencia de datos y bibliotecas-- para analizar 249,899 registros de una de las colecciones más importantes de Ciencias Sociales y Humanidades de América Latina: la del catálogo de la Biblioteca Daniel Cosío Villegas de El Colegio de México.

Metodología

A través del portal de analíticas del Grupo Ex Libris, se extrajeron los encabezamientos de materia de todos los 249,899 registros de libros de la colección de la Biblioteca Daniel Cosío Villegas. Los encabezamientos de materia fueron subdivididos a su vez en tres niveles a partir de los subencabezamientos, sin distinguir entre sus tipos –geográficos, cronológicos y de forma (ver Salta et al., 2015)– sino sólo tomando en cuenta su posición (primer subencabezamiento, segundo, etcétera). Por ejemplo, México--Historia--1821-1861 fue dividido en: México, Historia, 1821-1861.
Se estudió la relación entre temas utilizando técnicas de minería de reglas de asociación. Estas procuran descubrir implicaciones de la forma $I \rightarrow i$ donde I es un conjunto de objetos y i es un objeto en particular, ambos tomados de un universo de objetos, en este caso temas. El soporte de I se define como el número de registros para los cuales I es subconjunto. La confianza se define como el soporte de $I \cup i$ entre el soporte de I (Leskovec, 2010).

Se debe notar que la frecuencia de los temas asociados a los registros es sumamente baja como se puede observar en la Tabla 1, lo cual puede deberse a que, tratándose de una biblioteca especializada en ciencias sociales y humanidades los temas que se asignan son muy específicos, a fin de que el usuario especializado pueda encontrar lo que realmente le sirve.



	Percentiles					
Tema	25%	50%	75%	85%	95%	99%
1	1	1	3	5	22	129
2	1	1	3	6	27	219
3	1	1	3	6	28	170

Tabla 1

Asimismo, es de notar que 231,052 (92.45%) de los registros tienen un encabezamiento de materia; 152,414 (treinta por ciento menos) llega a tener dos encabezamientos de materia y sólo 29.89% tuvo tres. Por este motivo, los encabezamientos se concatenaron verticalmente para observar indistintamente las relaciones entre éstos. Se utilizó el algoritmo *a priori* y la elección de los umbrales se llevó a cabo de manera manual; se generaron 13 conjuntos de reglas de asociación con variaciones en los umbrales de confianza y soporte. Cada uno de estos conjuntos de reglas de asociación induce un grafo que se puede visualizar y explorar como se muestra más adelante. Umbrales demasiado permisivos inducen redes que tienen demasiadas relaciones como para poderse explorar manualmente y umbrales demasiado restrictivos inducen redes que no tienen suficientes relaciones como para poder decir algo interesante sobre la estructura de los datos en su totalidad. Finalmente se eligió una red que presenta un balance entre cantidad de información e interpretabilidad. El 'soporte' mínimo fue de 0.0001 (ver Gráfico 1) y la 'confianza' mínima de 0.4 (ver Gráfico 2) y la matriz de incidencia derivada de las reglas encontradas se utilizó para generar un grafo para la exploración visual del conjunto de asociaciones descubiertas. Para crear esta versión gráfica utilizamos la exportación de R a Gephi (Yon and Yon, 2015), la 'confianza' como un peso para los vértices y Fruchterman Reingold (1991) como algoritmo para el diseño. Dimos color a los nodos de acuerdo con su modularidad, es decir, de acuerdo a las "comunidades" de nodos que se crean por la fuerza de sus relaciones (Blondel et al, 2008). La alta modularidad de la red prueba lo conectados que están los nodos en sus grupos y lo desconectados que están de nodos fuera de su red.

Resultados

Como hemos mencionado antes, los encabezamientos fueron divididos en los subencabezamientos que los anidan. Retomando el ejemplo anterior: "México--Historia--1821-1861" fue codificado como:

- Subject 1.1 - México
- Subject 1.2 - Historia
- Subject 1.3 - 1821-186

Este modelado de los datos, fue pensado para permitir una cierta exploración "gramática" de la asignación temática. Es decir, que permitiera ver qué niveles "sintácticos" se relacionan en qué orden con otros niveles. En números, la red tiene 394 nodos (subencabezamientos) y 339 vértices (asociaciones). De los nodos, 203 son del primer nivel, 109 del segundo, 33 de la combinación de un encabezamiento del primer nivel con el tercero, y cuatro de la combinación del primer nivel con el cuarto. El total asociaciones o reglas de implicación (si encabezamiento I aparece también i) fue de 339. De éstas la mayoría ocurre sólo en 25 registros, es decir, tuvieron un soporte bajo (ver Gráfico 1). Sin embargo, esto no es tan poco considerando lo que hemos dicho antes de la naturaleza especializada de esta biblioteca. Por otro lado, las confianzas observadas presentan una distribución menos concentrada que la de los soportes (ver Gráfico 2).

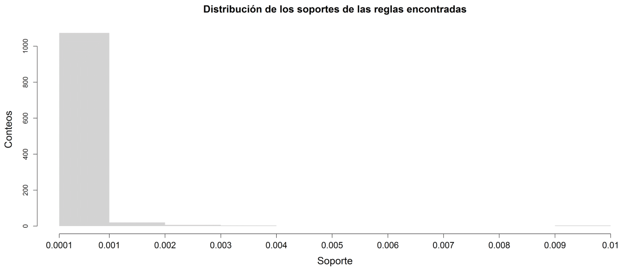


Gráfico 1

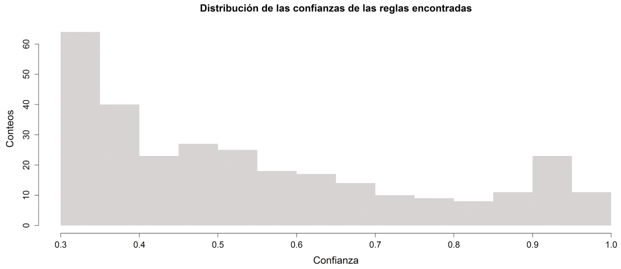
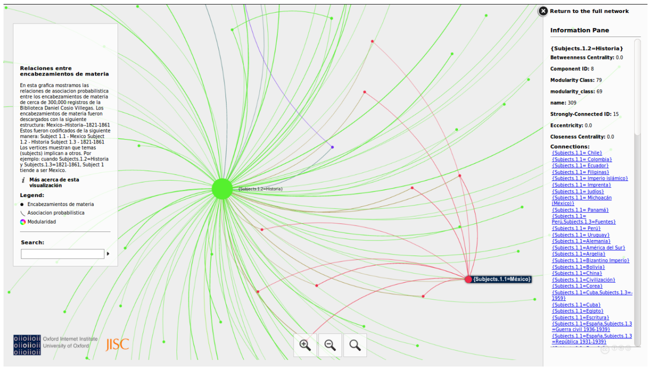


Gráfico 2

De la red de grafo interactiva que obtuvimos con el uso de Gephi y el *plug-in* de Sigma.js, pudimos identificar que el nodo con mayores asociaciones o reglas es 'Historia' en su posición como "Subject 1.2" y que entre sus asociaciones existen dos nodos de distinta modularidad y nivel (ambos "Subject 1.1"): 'México' (ver Imagen 1) y 'España' (ver Imagen 2).



Relaciones entre encabezamientos de materia

En esta grafica mostramos las relaciones de asociacion probabilistica entre los encabezamientos de materia de cerca de 300,000 registros de la Biblioteca Daniel Cosio Villegas. Los encabezamientos de materia fueron descargados con la siguiente estructura: Mexico--Historia--1821-1861 Estos fueron codificados de la siguiente manera: Subject 1.1 - Mexico Subject 1.2 - HIstoria Subject 1.3 - 1821-1861 Los vertices muestran que temas (subjects) implican a otros. Por ejemplo: cuando Subjects.1.2=Historia y Subjects.1.3=1821-1861, Subject 1 tiende a ser Mexico.

i **Más acerca de esta visualización**

Legend:
- ● Encabezamientos de materia
- Asociacion probabilistica
- Modularidad

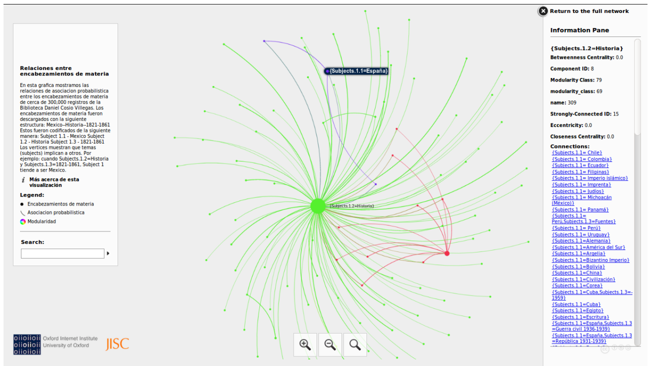
Search:

Return to the full network

Information Pane

{Subjects.1.2=Historia}
Betweenness Centrality: 0.0
Component ID: 8
Modularity Class: 79
modularity_class: 69
name: 309
Strongly-Connected ID: 15
Eccentricity: 0.0
Closeness Centrality: 0.0
Connections:
{Subjects.1.1= Chile}
{Subjects.1.1= Colombia}
{Subjects.1.1= Ecuador}
{Subjects.1.1= Filipinas}
{Subjects.1.1= Imperio islámico}
{Subjects.1.1= Imprenta}
{Subjects.1.1= Judíos}
{Subjects.1.1= Michoacán (México)}
{Subjects.1.1= Panamá}
{Subjects.1.1= Perú,Subjects.1.3=Fuentes}
{Subjects.1.1= Perú}
{Subjects.1.1= Uruguay}
{Subjects.1.1=Alemania}
{Subjects.1.1=América del Sur}
{Subjects.1.1=Argelia}
{Subjects.1.1=Bizantino Imperio}
{Subjects.1.1=Bolivia}
{Subjects.1.1=China}
{Subjects.1.1=Civilización}
{Subjects.1.1=Corea}
{Subjects.1.1=Cuba,Subjects.1.3=-1959}
{Subjects.1.1=Cuba}
{Subjects.1.1=Egipto}
{Subjects.1.1=Escritura}
{Subjects.1.1=España,Subjects.1.3=Guerra civil 1936-1939}
{Subjects.1.1=España,Subjects.1.3=República 1931-1939}

Imagen 1



Relaciones entre encabezamientos de materia

En esta grafica mostramos las relaciones de asociacion probabilistica entre los encabezamientos de materia de cerca de 300,000 registros de la Biblioteca Daniel Cosio Villegas. Los encabezamientos de materia fueron descargados con la siguiente estructura: Mexico--Historia--1821-1861 Estos fueron codificados de la siguiente manera: Subject 1.1 - Mexico Subject 1.2 - HIstoria Subject 1.3 - 1821-1861 Los vertices muestran que temas (subjects) implican a otros. Por ejemplo: cuando Subjects.1.2=Historia y Subjects.1.3=1821-1861, Subject 1 tiende a ser Mexico.

i **Más acerca de esta visualización**

Legend:
- ● Encabezamientos de materia
- Asociacion probabilistica
- Modularidad

Search:

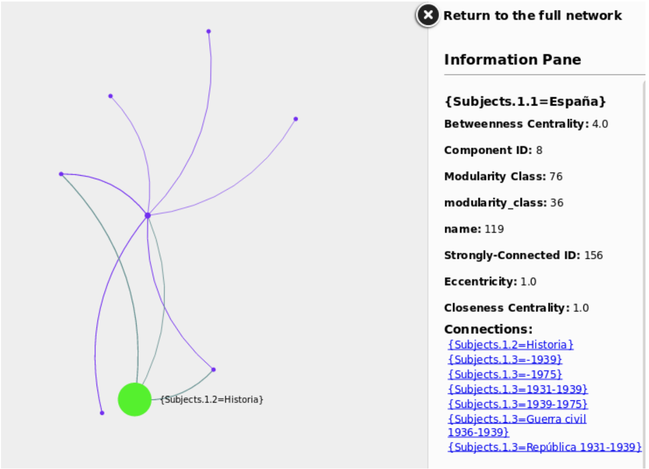
Return to the full network

Information Pane

{Subjects.1.2=Historia}
Betweenness Centrality: 0.0
Component ID: 8
Modularity Class: 79
modularity_class: 69
name: 309
Strongly-Connected ID: 15
Eccentricity: 0.0
Closeness Centrality: 0.0
Connections:
{Subjects.1.1= Chile}
{Subjects.1.1= Colombia}
{Subjects.1.1= Ecuador}
{Subjects.1.1= Filipinas}
{Subjects.1.1= Imperio islámico}
{Subjects.1.1= Imprenta}
{Subjects.1.1= Judíos}
{Subjects.1.1= Michoacán (México)}
{Subjects.1.1= Panamá}
{Subjects.1.1= Perú,Subjects.1.3=Fuentes}
{Subjects.1.1= Perú}
{Subjects.1.1= Uruguay}
{Subjects.1.1=Alemania}
{Subjects.1.1=América del Sur}
{Subjects.1.1=Argelia}
{Subjects.1.1=Bizantino Imperio}
{Subjects.1.1=Bolivia}
{Subjects.1.1=China}
{Subjects.1.1=Civilización}
{Subjects.1.1=Corea}
{Subjects.1.1=Cuba,Subjects.1.3=-1959}
{Subjects.1.1=Cuba}
{Subjects.1.1=Egipto}
{Subjects.1.1=Escritura}
{Subjects.1.1=España,Subjects.1.3=Guerra civil 1936-1939}
{Subjects.1.1=España,Subjects.1.3=República 1931-1939}

Imagen 2

A su vez, la plataforma permite explorar más a fondo el encabezamiento 'España' y darse cuenta, por ejemplo, de que este tema en primera posición tiene fuertes relaciones con subencabezamientos de la tercera dimensión que corresponden a los periodos históricos relevantes en la historia de ese país:



Return to the full network

Information Pane

{Subjects.1.1=España}
Betweenness Centrality: 4.0
Component ID: 8
Modularity Class: 76
modularity_class: 36
name: 119
Strongly-Connected ID: 156
Eccentricity: 1.0
Closeness Centrality: 1.0
Connections:
{Subjects.1.2=Historia}
{Subjects.1.3=-1939}
{Subjects.1.3=-1975}
{Subjects.1.3=1931-1939}
{Subjects.1.3=1939-1975}
{Subjects.1.3=Guerra civil 1936-1939}
{Subjects.1.3=República 1931-1939}

Imagen 3

En resumen, este tipo de exploración permite al usuario familiarizarse con las reglas "gramaticales" de la asignación temática pues puede "ver" tanto los niveles "sintácticos" de los temas como las formas en que se relaciona con otros, además de que incluye un botón de búsqueda de encabezamientos que permite interactuar de manera directa con el grafo (disponible en linea).

Reflexión final

Nosotros, como lo sugieren Nurmikko-Fuller et al., estamos conscientes de que si las bibliotecas quieren dar acceso a recursos de información relevantes para nuevas áreas de investigación, deben evolucionar a métodos más sofisticados y semánticos de asignación temática para proporcionar nuevos puntos de acceso que correspondan más al lenguaje natural y que permitan identificar las relaciones temáticas con mayor claridad.

Sin embargo, en lo que este paso puede ser dado en México y Latinoamérica, creemos que el uso de herramientas y métodos de las humanidades digitales pueden ayudar a analizar los datos generados en la organización de la información e incluso útil para la formación del catalogador, que aprende a asignar-elaborar los temas y con esta herramienta podría tener un acceso visual a la "sintaxis temática" de ciertos términos. En este mismo sentido, un acercamiento así, podría ser usado como elemento pedagógico de los cursos de investigación documental en el que los estudiantes deben aprender a familiarizarse con los lenguajes controlados. Otra aplicación de este trabajo, podría ser en la evaluación de colecciones para determinar las fortalezas y carencias temáticas, de acuerdo con la especialidad que la biblioteca declara. Análisis más detenidos pueden ayudarnos a determinar la representación cronológica, autoral, lingüística o geográfica de un acervo. En fin, consideramos que al continuar el análisis y desarrollo de este proyecto podremos aportar otro tipo de metodología no sólo para evaluar las colecciones sino para acercarse a ellas.

Bibliografía

Blondel, V., et al. (2008). "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, P1008.

Duguid, T. (2015), "BigDIVA: Big Data, Big Visuals, Big Searches, and Big Results." *Texas Digital Humanities Conference 2015.* University of Texas Arlington, Texas.

Fruchterman, T. M., & Reingold, E. M. (1991). *Graph drawing by force-directed placement. Software: Practice and experience*, 21(11), pp. 1129-64.

Green, R. *(2001). "Relationships in the o*rganization of knowledge: an overview." *Relationships in the organization of knowledge*. Springer Netherlands, pp. 3-18.

Nurmikko-Fuller, T., Jett, J., Cole, T., Maden, C., Page, K., Downie, J. (2016). "A Comparative Analysis of Bibliographic Ontologies: Implications for Digital Humanities". *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 639-42.

Leskovec, J., Rajaraman, A., Jeffrey, U. (2010). *Mining of Massive Datasets.* Cambridge University Press, U.K., pp. 205-14.

Salta, G., Cravero C., Saloj, G. (2005) "Lista de encabezamientos de materia de la Biblioteca del Congreso de los Estados Unidos: características generales". *Información, Cultura y Sociedad*, 12. pp. 85-97

Yon, G. V., & Yon, M. G. V. (2015). Package 'rgexf'.

Cäsar Flaischlen's "Graphische Litteratur-Tafel" – digitising a giant historical flowchart of foreign influences on German literature

Angelika Hechtl
angelika.hechtl@gmail.com
WU Wien, Austria

Ingo Börner
ingo.boerner@univie.ac.at
Universität Wien, Austria

Frank Fischer
ffischer@hse.ru
Higher School of Economics Moscow, Russia

Peer Trilcke
trilcke@uni-potsdam.de
Universität Potsdam, Germany

Introduction

By publishing his "Graphische Litteratur-Tafel" [Graphic Literature Table] in 1890, German writer Cäsar Flaischlen (1864–1920) aimed to portray the influences of foreign literatures on the development of German literature. Flaischlen produced a 58×86.5-cm poster, depicting German Literature as a stream with feeder rivers from mainly other European (national) literatures. His chart covers the development of German literature from its beginnings with various sources, forming two parallel rivers subsumed under the concepts of "Volkspoesie" [folk poetry] and "Kunstpoesie" [artistic poetry], intertwining and finally converging into one broad stream of German literature. The broadening river reaches from the beginnings of German literature at around 750 to Flaischlen's present, the 1890s.

Cäsar Flaischlen was not much of a practising literary scholar, nor an academic. After completing a dissertation around the same time as he published his "Graphische Litteratur-Tafel", he left academia to continue writing (dialect) poetry, novels and plays, while working as an editor for arts and literary magazines.

The flowchart, although being a flamboyantly beautiful one, is not as novel and unique, as one might think: It follows a long tradition of visualising developments along the axis of time (cf. Rosenberg and Grafton, 2010) and follows the patterns of the highly influential graphical visualisation of history, "Strom der Zeiten", published in 1804 by Austrian historiographer Friedrich Strass.

Being a representative of positive thinking of his time, Flaischlen builds on the time-stream metaphor, but strives to connect this idea with the exact sciences. Although he does not mention the sources used for compiling his chart, this early visualisation of literary history relies on some kind of data, even if, in his 8-column preface, Flaischlen deemphasises the connection between quantitative evidence and visualisation: For example, he points at the fact that the breadth of the stream was not calculated mathematically ("nicht mathematisch berechnet"), but nonetheless, there seems to be a connection between the selection and especially placement and typographical styling of influencing and influenced authors' names on the chart.

As this example illustrates, the information density of the chart is enormously high:

Flaischlen includes names of authors, texts, literary groups and schools and uses typography (font, font size, font decoration, colour), symbols (circles in various sizes, Roman and Latin numerals), shading of creeks and rivers and language-information to visualise the information.

Our digital edition of Cäsar Flaischlen's "Graphische Litteratur-Tafel" aims to make digitally available the information deeply encoded in the table. Extracted data points are presented in two shapes: on an easily navigable website and in a machine-readable version.

For this reason, the preface was OCRed and encoded in XML according to the TEI guidelines. The flowchart was scanned and transcribed following the recommendations on "Advanced Uses of <surface> and <zone>" (cf. TEI guidelines) to also record spatial information. Coordinates of authors and texts were calculated by help of the GIMP Image-Map editor, then linked to corresponding authority files (VIAF, GND, and Wikidata). The cartographic inventory of the map is marked-up and made searchable by using CSS as a descriptive language for capturing the rendering within TEI @style attributes.

The TEI data is provided via a GitHub repository and processed via XSLT for displaying the flowchart as a web page.

The interface combines the three separate sections of Flaischlen's map and allows for searches via an index. For presentation on the web, the sections were stitched and put together into one giant river. In an analogue format, the river would be almost three meters long, now one can scroll seamlessly down and up the river online. A prototype of the edition is available on the project website.

Flaischlen's map is not only an inspiring prototype for contemporary attempts to visualise data of literary history with "graphs, maps and trees" (Moretti, 2007), but also challenges the development and encoding of older graphical representations of (literary) history on a methodological level.

Flaischlen's 1890 visualisation "Graphische Litteratur-Tafel" deserves further recognition and our online edition could serve as a test case for other – so far unknown and/or not edited – flowcharts of literary history.

Bibliography

Börner, I., Fischer, F., Hechtl, A. and Trilcke, P. (eds.) (2016). *Cäsar Flaischlen's 'Graphische Litteratur-Tafel'. A Digital Edition.* (Version: Alpha, October 2016). Available at: http://litteratur-tafel.weltliteratur.net.

Flaischlen, C. (1890). *Graphische Litteratur-Tafel. Die deutsche Litteratur und der Einfluß fremder Litteraturen auf ihren Verlauf vom Beginn der schriftlichen Ueberlieferung an bis heute in graphischer Darstellung.* Berlin: Behr's Verlag.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. London and New York: Verso.

Rosenberg, D. and Grafton, A. (2010). *Cartographies of Time*. New York: Princeton Architectural Press.

Strass, F. (1803). *Der Strom Der Zeiten*. [online]. Available at: http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~281767~90054624.

TEI Consortium (2016). *TEI P5 – Guidelines for Electronic Text Encoding and Interchange. Version 3.0.0.* Available at: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

German History–*Digital*: A Platform for Transnational Historical Knowledge Co-creation

Matthew Hiebert
hiebert@ghi-dc.org
German Historical Institute (Washington DC)
United States of America

Simone Lässig
laessigs@ghi-dc.org
German Historical Institute (Washington DC)
United States of America

Andreas Witt
andreas.witt@uni-koeln.de
University of Cologne, Germany

The German Historical Institute Washington (GHI) is in the development phase of German History-*Digital* (GH-D), a transatlantic digital initiative to meet the scholarly needs of historians and their students facing new historiographical and technological challenges. In the proposed paper we will discuss the research goals, methodology, prototyping, and development strategy of GH-D as infrastructure to facilitate transnational historical knowledge co-creation for the large community of researchers and students already relying on digital resources of the GHI and for the growing constituency of citizen scholars.

Despite its great progress, the digital humanities have yet to broadly impact research in German history. The past ten years have witnessed the proliferation of online resources relevant to the field, yet these materials largely remain siloed in different systems, with material difficult to discover or gather by scientists into corpora. Historians themselves are today increasingly producing scholarly content in digital form, but there remains no established criteria for the peer review of digital publications and projects. This ultimately limits the time and energy the research community is willing to invest into digital knowledge production and thus confines the Digital Humanities' potential for growth.

Preservation and future access to digital materials is also of critical importance, particularly in the North American context where continental and national digital research infrastructures for digital humanities are lacking; there is currently no equivalent to European research infrastructures like CLARIN or DARIAH. In respect to scientific methodology, there is growing expectation that historians take advantage of an abundance of digital tools, yet there remains insufficient integration between tools for historical research and between tool sets and online resources. The importance of citizen science and knowledge co-creation for the future of historical research is also recognized, yet for these developments to occur there must exist beyond e-lists and other legacy communication technologies scholarly environments for the creation of area-specific research communities, scientific collaboration, and public engagement.

The planning for GH-D involved surveying over four hundred scholars of the many thousands already using digital resources produced by the GHI. The most prominent of these resources is the digital source collection „German History in Documents and Images" (GHDI), which is widely used at universities in the German and English speaking world. Launched in 2003 and currently undergoing a technical and conceptual revamp in conjunction with GH-D, GHDI currently includes thousands of pages of English-language translations of German historical texts, as well as images and maps, all of which are accessed by approximately 5,000 visitors per day. Our planning for GH-D also continues to involve consultations and workshops with expert historians and digital humanists, and the establishment of partnerships with institutions and major initiatives that share our concern for the future of history in the digital age.

The German History-*Digital* platform addresses needs of digital scholarship through five goals and integrated work packages concerted to these goals: *discovery, analysis, production, preservation, and community*.

We believe GH-D provides a new model in the design and development of a social knowledge creation environment for humanities-oriented research. The proposed paper will be structured by providing technical and theoretical explication of the core work packages within relevant DH contexts.

Discovery

A major challenge facing scholarship online is that a vast number of digital resources, particularly those produced independently by scholars or smaller institutions, do not have standardized metadata records and are not accessible via any centralized scientific index. GH-D involves development of a peer-reviewed index of scholarly digital objects using Dublin Core (DC) and CLARIN's Component MetaData Infrastructure (CMDI) standards via a customized Blacklight technology stack.

Preservation

For scholars developing historical digital projects in North America, there exists no inter-institutional infrastructure for preserving their data and making it openly available. With consultative and knowledge support from CLARIN-D, which is part of the European research infrastructure CLARIN, the GH-D project will establish the first portal to CLARIN in North America at GHI Washington. Central to this process is the implementation of a repository that allows a sustainable storage of the content and the inclusion in a digital environment to ease access, search and an interoperable data formats. The content of the repository and the repository itself adheres to international, widely accepted and supported standards. The high quality of the technical solution and the conformance to standards is secured by an independent organization that gives out the Data Seal of Approval. Like the majority of CLARIN centres in Germany, the GH-D will use a Fedora Commons repository with Apache Solr for indexing and search, components included in the technology stack of Project Hydra. Our partnership with CLARIN promotes open access, open science and knowledge co-creation in the North American context, and is an important component in the overall digital humanities research strategy of the GHI. As an institute of the Max Weber Stiftung, we are also in partnership with DARIAH-DE and arrangements have been made for DARIAH to provide long term preservation of GHI digital projects in their entirety, beginning with the first edition of German History in Documents and Images. Beginning with the GHDI project, GH-D is part of the DARIAH-DE Service Lifecycle program.

Production and Publication

As a knowledge co-creation platform, GH-D will bring together editors, researchers and citizen scientists in the development of innovative online projects. Three such pilot projects are currently in development based on customization, including support for TEI, and internationalization of the Scalar 2.0 platform. GH-D is using Scalar 2.0 for the baseline content management system, particularly on account of its interface features, support for RDF, connectivity to external repositories, Dublin Core support. Hypothes.is integration, and its multiple path navigation system.

Analysis

Historians are increasingly using digital humanities tools to analyze data and express their research findings. A further advantage of storing digital objects within the CLARIN repository the GHI wants to built up is that the full range of corpus linguistic analytic tools of CLARIN can be applied by scientists to GHI textual content. During the first phase of the project we also look to prototyping connectivity to PARTHENOS, another major European infrastructure project. PARTHENOS integrates within a virtual research environment (VRE) access to data from numerous national archives and a broad set of digital tools which can be chained together into analytic processing workflows.

Community

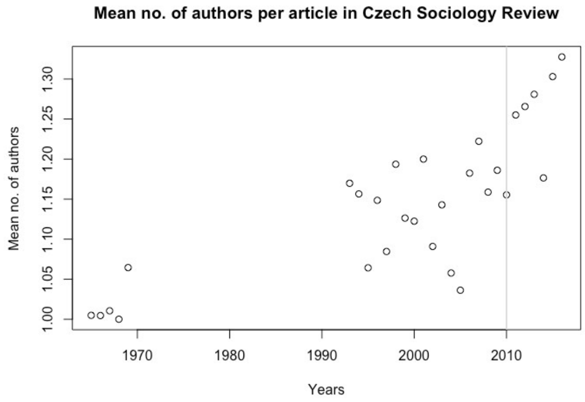
The GH-D platform integrates blog aggregation, an advanced discussion system, community-oriented tools, and social media, to facilitate collaborative knowledge communities and open research. This is a pioneering aspect of our project that will investigate the adoption by historians of social and community digital tools in their research activities. We also intend to make use of the unique role the GHI plays as a hub of transatlantic scholarly dialogue and a major knot within an international network of historians in order to facilitate connections between different scholarly communities.

Topics and genre changes in Czech sociological articles

Radim Hládik
radim.hladik@fulbrightmail.org
Czech Academy of Sciences, Czech Republic

Introduction

Since Kuhn's distinction between "textbook" science and "article" science, studies of scientific texts have established the social nature of academic writing. Analysis of scientific texts can yield epistemological, disciplinary, and historical insights. These texts are the arena in which knowledge claims are raised (Myers 1985), trials of strength held (Latour 1987), intradisciplinary boundaries drawn (Wolfe 1990), disciplinary histories traced (Bazerman 1988). Genres of scientific texts have thus been shown as socially enacted structures (Berkenkotter, Huckin, Ackerman 1994) rather than as transparent styles. Despite their many merits, the hitherto available empirical studies on scientific writing have been constrained by either a focus on early history (such as the original Philosophical Transactions of the Royal Society) or reductive sampling for traditional content analysis that allow researchers to grasp the otherwise immense textual data. Approaches inspired by digital humanities approaches offer new possibilities of studying disciplinary formations, as was demonstrated by Goldstone and Underwood (2014) in their distant reading of literary studies texts. This paper follows suit in reporting the results of applying digital humanities methods of text analysis to the corpus of research articles in sociology, specifically, in Czech Sociological Review. Writing in wide-scope disciplines, such as sociology, is of particular interest because it embodies the conflict between literature and the notion of social science (Lepenies 1988). Sociology has been revealed as a discipline of two writing cultures, monographic and journal (Wolfe 1990). The writing in sociology is also oscillating between the aspiration to the positivist ideal of science (Leenhardt 1992) and the acceptance of diverse styles (Agger 2002). Abbot and Barman (1997) have concluded, on the basis of sequence comparison, that research articles in sociology lack "rhetorical rigidity". The discipline thus offers a particularly opportune resource for the analysis of genre and topical variations. Czech Sociological Review was chosen as an example of a "core" journal in the country. As Oromaner (2008) demonstrated, "core journals" in sociology have tendency to become central to the discipline's "intellectual integration". Thus the results of the analysis can be taken as indicative of the mainstream tendencies in Czech sociology. The focus on the Czech sociology has the additional advantage of representing a interesting example of a discipline undergoing substantial transformation in the wake of academic and wider societal changes that came about with the fall of the Communist Party regimes in 1989. Also, the journal offers open access to its content. The data for the analysis were scrapped from the journal's website in September 2016 and the resulting data set contains 3483 articles. A preliminary exploration of metadata revealed noteworthy patterns around the year in which a new policy for science evaluation had been introduced (cf. the figure).



The data collection and analysis is carried out using R language. Besides crude measures of the corpus, the paper will also report the analysis of the textual data, using text mining techniques to comment on the issues that have been raised in the available literature. Topic modeling through LDA model will be used to assess the topical changes across time. Annual frequencies of particular words will be used as indices of changes of the transforming disciplines (this includes, especially, the words relating contemporary sociology to its "communist" variety, such as references to Marx or "communism"). Multidimensional scaling will then be employed to reveal term clustering around further keywords that arguably important in sociology. Quantitative bias, or a lack of thereof, will be measured by the presence of numbers. The overall purpose of the analysis is to address the questions raised in pre-existing literature using the specific example of a Czech social science discipline and to demonstrate the usefulness of text mining techniques in the analysis of scientific writing.

Bibliography

Abbott, A., and Barman, E. (1997). "Sequence Comparison Via Alignment and Gibbs Sampling: A Formal Analysis of the Emergence of the Modern Sociological Article." *Sociological Methodology* 27 (1): 47–87.

Agger, B. (2002). "Sociological Writing in the Wake of Postmodernism." *Cultural Studies / Critical Methodologies* 2 (4): 427–59.

Bazerman, C. (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science.* Madison: University of Wisconsin Press.

Berkenkotter, C., Huckin, T. N., and Ackerman, J. (1994). “Social Context and Socially Constructed Texts.” In *Landmark Essays on Writing across the Curriculum*, edited by Charles Bazerman and David R. Russell. London, New York: Routledge.

Goldstone, A., and Underwood, T. (2014). “The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us.” *New Literary History* 45 (3): 359–84.

Leenhardt, J. (1992). “Writing and ‘Scientific Discourse’ in Sociology.” *History of the Human Sciences* 5 (1): 63–71.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society.* Cambridge, Mass: Harvard University Press.

Lepenies, W. (1988). *Between Literature and Science: The Rise of Sociology. Ideas in Context.* Cambridge [Cambridgeshire]; New York: Cambridge University Press.

Wolfe, A. (1990). “Books vs. Articles: Two Ways of Publishing Sociology.” *Sociological Forum* 5 (3): 477–89.

Oromaner, M. (2008). “Intellectual Integration and Articles in *Core Sociology Journals*, 1960–2000.” *American Sociologist* 39 (4): 279–89.

RAT 2.0

Winfried Höhn
winfried.hoehn@uni.lu
ILIAS Lab, University of Luxembourg, Luxembourg

Christoph Schommer
christoph.schommer@uni.lu
ILIAS Lab, University of Luxembourg, Luxembourg

Historical maps are progressively digitized and added to the inventory of digital libraries. Beside their value as historical objects, such maps are an important source of information for researchers in various scientific disciplines. This ranges from the actual history of cartography and general history to the geographic and social sciences. However, for most of these digital libraries, the available metadata include only limited information about the content of the maps, for example author, title, size, and/or creation date.

Whereas given information extraction methods are designed for modern maps and mostly limited to certain types that share similar graphical features, there exist a limited number of tools that rely on a manual recording to visualize certain properties such as distortions as well as support a content-based querying. Examples concern the development of places over time, toponym changes over time, and the identification of the position of places (historical map vs. modern map). This also applies to place markers and text labels, which contain inherent information and so the annotation and geo-referencing of place markers is a crucial task, which can be supported with computer based tools (Budig and Dijk, 2015, Höhn et al., 2013, Shaw and Bajcsy, 2011, Simon et al., 2011).

As already presented in previous contributions (Höhn and Schommer, 2016, Höhn et al., 2013), the Referencing and Annotation Tool RAT supports an identification of place markers in digitized historical maps. RAT facilitates a geo-referencing by suggesting the most likely modern places based on an estimated mapping. The suggestions can be constrained by additional filters, for example by applying a phonetic search (with the Kölner Phonetik) to places, which sound similar to names given on the map. This allows an identification of modern places, whose historic name has changed over time but where its name still is close. RAT performs a template matching algorithm based on the normalized cross-correlation for the identification of place markers. If there are colored place markers in a map, a color segmentation methodology can be used to detect these markers. With respect to the geo-referencing, RAT uses the implemented phonetic search and an estimation of the positions of the place markers.

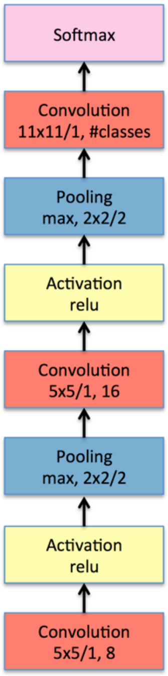


Figure 1. Architecture of the Convolutional Neural Network, which shows the operations used for processing the image

In addition to the original template-based place marker recognition algorithm, we integrated a place marker recognition algorithm based on convolutional neural networks (CNNs; RAT 2.0).

For these algorithms, the user is asked to manually annotate a small subset of the map. Regarding the template-based variant, the user has then to select a template for each type of place marker. From the templates, which are manually chosen by the user, the system creates automatically variations; based on the performance – measured with the annotated small subset of the map –, the best performing template variants are then chosen. In the template-matching algorithm, however, the normalized cross-correlation is used as a similarity measure because of its robustness against changes in brightness and contrast.

Regarding the detection trough a convolutional neural network, the manually annotated (small) subset of the map is used and split in a training and validation part. This is used to train the network, which has – at this stage – a very basic structure. It consists only of convolutional and pooling layers as presented in Figure 1.

Both the template matching approach and the convolutional neural network approach share similar performances. Our tests have shown – for the template matching approach – a detection precision of 98.2% and a recall rate (discovered place markers divided by all existing place markers on the map) of 87.7%. The convolutional neural network approach reaches only a precision of 94.4%, but gives a recall rate of 96.2%. So, there are more place markers found; but the result contains also some more wrong matches in between. Therefore, it depends on the use case, which result is "better", but for the manual post-correction it seems easier to check the CNN results for those additional wrong matches then finding the missed matches from the template-based approach.

The reason behind the use of the convolutional neural network approach has been an algorithmic limitation of the template matching approach. So far, RAT 2.0 uses only a fundamental convolutional system (at present, there are no additional techniques used, like, for example, data augmentation or pre-training).

As a future point, we work on training the convolutional neural network on multiple maps in order to find a classification model that learns the characteristics of place markers and that detects these on unseen maps.

Bibliography

Budig, B. and van Dijk, T. C. (2015). "Active Learning for Classifying Template Matches in Historical Maps." *International Conference on Discovery Science*, pp. 33-47.

Höhn, W., Schmidt, H.-G. and Schöneberg, H. (2013). "Semiautomatic Recognition and Georeferencing of Places in Early Maps." *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 335–38.

Höhn, W. and Schommer, C. (2016). "Annotating and Georeferencing of Digitized Early Maps." *Digital Humanities 2016*, pp. 807–808

Höhn, W. and Schommer, C. (2016). "RAT: A Referencing and Annotation Tool for Digitized Early Maps." *Digital Humanities BeneLux 2016*

Shaw, T. and Bajcsy, P. (2011). "Automation of Digital Historical Map Analyses." *Proceedings of the IS&T/SPIE Electronic Imaging*, Vol. 7869.

Simon, R., Haslhofer, B., Robitza, W. and Momeni, E. (2011). "Semantically Augmented Annotations in Digitized Map Collections." *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, pp. 199–202.

Georeferencing of Place Markers in Digitized Early Maps by Using Similar Maps as Data Source

Winfried Höhn
winfried.hoehn@uni.lu
ILIAS Lab, University of Luxembourg, Luxembourg

Christoph Schommer
christoph.schommer@uni.lu
ILIAS Lab, University of Luxembourg, Luxembourg

Early maps are usually only accessible for a small group of researchers and librarians because they are precious and fragile. In the age of Digital Humanities, online access and search in digitized historical documents and early maps allows people from all over the world to work with such artefacts of cultural heritage. However, the digitization solely generates images of the artefacts without any access to the semantics of the documents. For most digital libraries of early maps (e.g. Old Maps Online) the available metadata include only information about the map, e.g. author, title, size, creation date, covered region. Unfortunately, there is only little information about the data contained in the map. Thus, even if data about place development or toponym changes is present in the maps it is not easily accessible. Since a single map can easily contain many thousands of place markers, proper tool support and automation of the annotation and georeferencing of each single place marker are of interest.

For modern maps or aerial photos it is possible to use GIS software to georectify the images by specifying a few control points, thus this problem is seen as solved. But early maps contain many sources of distortion, for example inaccuracies during surveying, combining data from different sources, focusing on creating a visually pleasing map instead of an accurate one. So there is in general no simple mapping between modern geocoordinates and an early map. Our existing Referencing and Annotation Tool (RAT) (Höhn et al., 2013) already simplifies the annotation and georeferencing of place markers. RAT supports the annotation and georeferencing by using template matching to identify place markers and by suggesting the most likely modern places based on an estimated mapping between the pixel-coordinates and geocoordinates of the already georeferenced place markers. To further refine the suggestions a phonetic search can be used, where the historic spelling can be used to restrict the results to similarly sounding place names.

Even with tool support like provided by RAT the georeferencing and annotation process starts from scratch for each map. Despite the automation there is still manual effort needed for place marker annotation. Since early maps have often been copied from each other or share some underlying survey data, there should be some regularity between maps that we can exploit. To take advantage of the possible similarities in early maps we present an algorithm to identify similar maps and create a link between the place markers of these maps. This results in georeferencing an early map in relation to another early map, which can be much simpler than georeferencing in respect to modern data. When the maps are based on the same data, they share some of their distortions and so the transformation between them gets simpler. They will also more likely contain a similar set of places. This reduces the problem of identifying a matching place compared to a modern database containing all known places, even the smallest ones which will not be shown in medium or small scale maps.

Before we can apply the algorithm to a pair of maps we need to identify suitable maps. These are maps that already have some georeferenced place markers and share at least four mappings to modern places. Also the place markers in these maps must be already recognized, but not necessarily georeferenced.

The algorithm for linking corresponding place markers of two maps *A* and *B* can be split in two steps:

1. Estimation of a transformation between the maps.The coordinate mappings in **M**, a bidirectional mapping containing the coordinates of the matching place markers in the two maps, are used to calculate the projective transformation between map *A* and map *B*.
2. Extending the linked place markers. Using the projective transformation calculated in step 1, map *B* is transformed into the coordinate system of map *A*. We will refer to the transformed map *B* as map *B'*. For each place marker in *A* the nearest place marker from *B'* is located. If for the place marker from *B'* also the place marker from A is the nearest, following checks are done:
   - The second nearest place markers have to be at least two times further away than the distance of the two place markers under consideration.
   - Both place markers must be connected to some place marker contained in **M** through edges in a Delaunay-Triangulation (Lee and Schachter, 1980) of map *A* and map *B*.

   If both previous conditions are true, add the place markers to **M'**.

   If **M'** has more elements than **M**, then set **M** to **M'** and continue with step 1.

M' is the resulting correspondence between the two maps. The steps of the algorithm are visualized in Fig. 3. The right column corresponds to step 1 and the left one to step 2. The rows show the different iterations of the algorithm. For all examples, the following maps are used: "*Nova Franconiae descriptio/Sculptum apud Abrahamum Goos. - Amsterdam: Joannes Janßonius, 1626*" referred to as Goos and "*Franckenlandt = Francia orientalis/Per Gerardum Mercatorem – o.O., ca. 1600*" referred to as Mercator.

These example maps both contain about 900 place markers and an overlapping area with about 800 place markers. For Goos, all place marker locations have been manually verified and for Mercator, the result of the template matching was kept. This resulted in an automatic detection of the correspondence of 755 place markers between the maps.

Another use case of this mapping is, that we can compare the automatically found place markers from two maps. We can highlight the differences between the sets of automatically detected place markers from two maps. This allows easily investigating the differences in the two sets of manually or automatically identified place markers. The identified differences highlight specific areas in these maps for further investigation. Two examples for detected differences between similar maps are shown in Fig. 1 and 2.

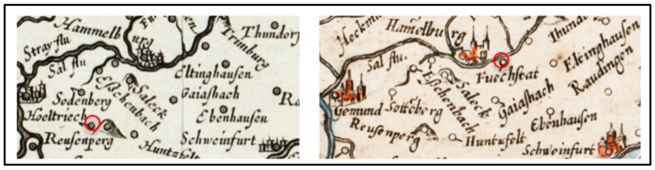


Figure 1. Corresponding map sections from Goos (left) and Mercator (right), where Goos has one place marker for Hoeltriech and Mercator one for Fuechstat which are both not in the other map.

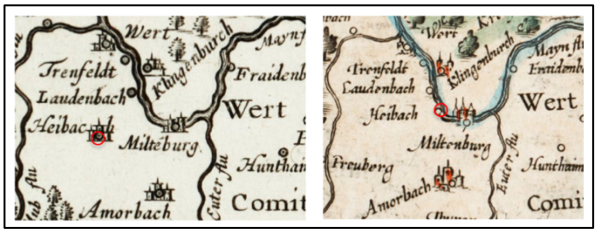


Figure 2. Corresponding map sections from Goos (left) and Mercator (right), where Heibach is in Mercator located at the river and in Goos far away from the river.

There can be various reasons for differences between two maps. First, there could have been an error in the detection of the place markers, which then helps to spot such problems. Second, it is a genuine difference between the two maps, which itself can have many reasons, e.g. different decisions which places should be included on a map or errors while placing the places on a map.

This work shows that it's possible to create a correspondence between place markers in different maps with not more effort than for georeferencing a map, which then only provides the region covered by the map. A similar map can be identified, if one exists in the database, and the place markers between the maps can be connected. This then allows reusing the georeferencing of single place markers from one map in the other map and identifying differences in the sets of place markers.

Future Work

This method can also be used to quickly identify similar maps and the differences in them. In this way it could be useful for researchers who want to find out which sources were used to create a map or who copied from whom. MapAnalyst (Jenny and Hurni, 2011) is an already existing tool for this purpose. If one map is considered as a possible copy of another map, MapAnalyst is a tool used by researchers to explore if this is true. The method proposed in this work would allow doing this kind of analysis on a larger scale while also highlighting the differences between the maps. Although this area was not our primary focus, we plan to evaluate the usefulness of our method on this task.

The information from linked place markers could help in further analysis of other metadata items, such as place type or place name. The linked place markers already make this information available from the other maps and it could for example be used to improve the OCR process of place names.

Appendix

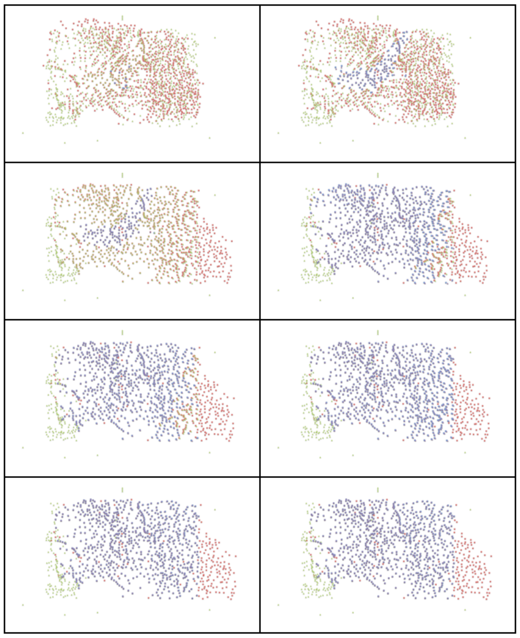


Figure 3: Visualization of the different steps in the matching procedure for Goos and Mercator, where the Mercator place markers are transformed into the Goos coordinate system. Triangles represent Goos place markers and squares Mercator ones. Green and red points don't have a mapping to a place marker in the other map, blue and purple ones have mappings. For further explanations see algorithm in main text.

Bibliography

Höhn, W., Schmidt, H.-G. and Schöneberg, H. (2013). "Semiautomatic Recognition and Georeferencing of Places in Early Maps." *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pp. 335–38.

Lee, D. T. and Schachter, B. J. (1980). "Two algorithms for constructing a Delaunay triangulation." *International Journal of Computer & Information Sciences*, 9(3), pp. 219–42.

Jenny, B., and Hurni, L. (2011). "Studying cartographic heritage: Analysis and visualization of geometric distortions." *Computers & Graphics*, 35(2), pp. 402–11.

Digital Access as an Equity Issue: The Community College and the Digital Divide

Polly Ruth Hoover
phoover@ccc.edu
Wright College, United States of America

In her nuanced discussion of teenagers and their use of digital technology, danah boyd argues that there are far-reaching consequences for assuming that young students are naturally digitally savvy and older adults are digitally hampered; we assume younger students are digital natives adept at using and understanding technology, while adults are digital immigrants, still learning the language and culture of the digital world (2014: 176-193). In fact, "Familiarity with the latest gadgets or services is often less important than possessing the critical knowledge to engage productively with networked situations, including the ability to control how personal information flows and how to look for and interpret accessible information" (boyd, 2014: 180), and she further cautions that "access to technology should not be conflated with use" (boyd, 2014:192).

But access is not universal. Even among teenagers, there are different levels of participation because of the access to the technology, the quality of the access related to the socioeconomic status of the teenager and the consequent different levels of digital skills (See Hargittai, 2008). While young, wealthier students may have access without sufficient knowledge about the limitations and challenges of their digital access, students from lower socioeconomic groups may not even have access except through the technology provided by their schools, libraries and other local institutions. The economic divide contributes to the digital divide.

This digital divide is further exacerbated by the institutional divisions and economic resources between two-year comprehensive community colleges and four-year institutions. Anne B. McGrail, a community college instructor herself, in her discussion of the role digital humanities might play in the community colleges, asks

"What are community colleges and their students missing out on in their exclusion from DH discourse, and how might intentional engagement with DH methods and tools help community college students become active agents of discovery and change in their lives?" (McGrail, 2016). These questions are particularly pointed given that many of the students at the comprehensive community colleges intend to transfer to four-year institutions and see the two-year colleges as an affordable first entry into college. If digital humanities is "the next big thing" as William Pannapacker announced in his 2010 blog in The Chronicle of Higher Education, is digital access another inequity in the system of higher education?

But how are digital access and inequality linked? At traditional four-year institutions, instructors may find the digital divide initially a challenge because of the access that students have had prior to their matriculation, but once they are enrolled, the institution provides digital access and support for both students and instructors. The learning curve may be steep for some students, but there are footholds on the way up. At the community colleges, however, which typically enrolls a variety of diverse students, a range from nontraditional students, who may be returning students apprehensive about their abilities, who may be limited in time and financial resources, and who are less likely to have access to or to use digital technology, to the more traditional students who nevertheless do not have the economic resources to access the newest digital resources, the digital divide begins at enrollment and may continue after enrollment. Both groups may be under-prepared, under-resourced and have issues of persistence. Yet, contrary to the four-year institutional model, the divide is further complicated because community colleges and their instructors, many of whom are adjuncts with peripatetic careers, are also under-resourced and lacking strong support for the digital humanities. The digital divide begins at enrollment and continues even after enrollment.

This presentation directly addresses these issues of digital access and equity among community college students. In particular, I examine the implications of the digital divide among the community college population: the problems of access and use among community college students both before and after they enroll, the different levels and aspects of digital preparedness for students, instructors and administrators, and the institutional issues of support and resources to create robust digital humanities at the comprehensive community college. I discuss in detail academic planning and creation of institutional support including establishment of Makerspaces and digital labs, professional development opportunities for faculty, and the creation of assignments that address some of these issues.

Bibliography

boyd, d. (2014). It's Complicated. New Haven. 176-193

Hargittai, E., (2008) "The Digital Reproduction on Inequality," in Grusky, Social Stratification: Class, Race, and Gender in Sociological Practice, Boulder, CO. 936-944

McGrail, A. B. (2016). "The 'Whole Game': Digital Humanities at Community Colleges," in Gold and Klein, eds., Debates in the Digital Humanities 2016, Minneapolis. 17.

Measuring Canonicity: a Network Analysis Approach to Poetry Anthologies

Natalie M. Houston
natalie_houston@uml.edu
University of Massachusetts - Lowell
United States of America

Pierre Bourdieu has theorized culture as a field, a space constituted by the structural hierarchies and interactions among people and institutions. These relations encompass social and economic positions; competitive and cooperative intentions; as well as what he calls "position-takings"— all of the actions or decisions that produce a work of art and its cultural value. In order to understand the history of literature, Bourdieu argues that we must examine how its meaning and cultural significance gets produced within the cultural field: "the sociology of art and literature has to take as its object not only the material production but also the symbolic production of the work, i.e. the production of the value of the work" (Bourdieu, "Field" 37). Anthologies are a central mechanism in the symbolic production of artistic value as they present a selective literary canon as representative of a defined field of study.

Anthologies of literature thus both reflect and contribute to the structure of the cultural field, whether by reproducing its hierarchies of value or by contesting them. Anthologies frequently reprint some number of texts that have already been anthologized, along with new selections that may themselves over time become canonical. The changing selections presented in literary anthologies over time offer a microcosm of the field of cultural production as modeled by Bourdieu, a "force-field acting on all those who enter it . . . in a differential manner according to the position they occupy there" (Bourdieu, Rules 232). Each decision by an editor about which authors and texts to include in an anthology inevitably responds in some way to the decisions and values previously circulated in the cultural field. Anthologies are an important mechanism by which value gets assigned to particular texts in relation to one another: "Canonicity is not a property of the work itself but of its transmission, its relation to other works in a collocation of works" (Guillory 55). Thus to understand canonicity we must examine the relationships among literary works within the mode of their transmission, the anthology. In this paper I adapt methods of network analysis to examine the structure of the relationships among the poems included in 30 anthologies of Victorian literature published from 1880-2002.

Rationale

Literary canon formation and change has long been of concern to literary scholars. Writing in 1979, in terms that prefigure our current concerns with the scope of computational analysis, Alastair Fowler admits, “The literature we criticize and theorize about is never the whole. At most we talk about sizable subsets of the writers and works of the past. This limited field is the current literary canon” (97). Fowler defined six types of literary canons arising from various constraints and choices: official, personal, potential, accessible, selective, and critical. Anthologies are frequently taken to represent the literary canon because in themselves they constitute what Fowler calls a “selective canon,” a subset of works chosen by an editor (or editorial team) presumably for particular reasons. Anthologies make works accessible to a wide range of readers and also contribute to (or potentially constrain) the pedagogical canon, the subset of works that are taught (Harris 112-13). Anthologies often reflect the critical canon, as works that become interesting to scholars gradually start showing up in anthologies.

In the so-called “culture wars” of the 1980s-1990s, debates about multiculturalism and changes to the humanities curriculum frequently focused on anthologies and syllabi as synecdoches for university education (Graff). As John Guillory suggests, both the conservative and progressive sides in this debate tended to rely on “an ideology of tradition” which invokes “an autonomous history of literature, which is always a history of writers and not of writing” (Guillory 63). For Guillory and Bourdieu, the history of writing can only be understood in relation to mechanisms of cultural value.

When literary scholars write about anthologies, they frequently describe how the selective canon changes over time; examine the ideological assumptions that undergird the selection process; and point out historical or thematic gaps in anthology coverage. Yet the method that they use for doing this primarily rests upon counting authors (Golding, Harris, Latane). Others use the number of pages allotted to each author as a proxy for importance or weight within the anthology (Bode, Damrosch, Lecker).

Such approaches reify the ideology of tradition in assuming that author names alone adequately describe the complexities of literary history. As Guillory suggests, “histories of canon formation, when they consist primarily of a narrative of reputations, of the names which pass in and out of literary anthologies explain nothing. Such narrative histories fail to recognize generic or linguistic shifts which underlie the fortunes of individual authors by establishing what counts as literature at a given historical moment” (Guillory 64). By offering a network analysis approach that allows us to explore “what counts as literature” rather than just “who counted” in different anthologies, we can better understand the structures of value instantiated and reproduced in these collections and thereby better understand the history of taste and value. For example, rather than simply noting the unsurprising fact that the Victorian poet laureate Alfred Tennyson is included in all anthologies of Victorian literature and accepting that he is thereby a canonical figure, this approach allows for discovering more precisely which poems by Alfred Tennyson were valued in 1880, 1930, or 1980, and how those choices corresponded with selections from other poets.

Method

In this paper I explore the utility of several different approaches to analyzing the structures of canon formation as represented in poetry anthologies. These include:

- examining the structure of the bimodal network created between anthologies and poems;
- measuring anthology similarity based on textual couplings;
- examining the co-printing network created by connecting each poem in an anthology to every other poem printed in that same anthology
- Each of these approaches will be explored using the full dataset and by using chronological slices that will further the understanding of historical changes in the anthology canon.

In the first approach, I treat the relationships of poems to anthologies as a bimodal affiliate network. Although some researchers avoid bimodal representations of relationships because standard measures of centrality and other metrics do not apply, force-directed visualization of bimodal data “is often extremely effective for transmitting a holistic understanding of the whole dataset” (Borgatti 10). Faust and Borgatti have each recommended approaches to calculating centrality for affiliation networks that I will also explore. I also examine centralization (core-periphery structure) and structural equivalence measures for the affiliation network.

The remaining two approaches derive from established practices in bibliographic network analysis, particularly bibliographic coupling and co-citation analysis. Although co-citation analysis has largely overtaken bibliographic coupling in recent decades, recent comparative studies suggest that the utility of each approach varies depending on the timeframe and construction of the citation dataset (Boyack et al).

Bibliographic coupling draws an edge between two documents which each cite a third (Kessler). To understand the similarities among these anthologies, I create a textual coupling network, which consists only of anthology nodes, and create an edge between each anthology that prints the same poem. This network shows the degree of similarity between anthologies and filtering the poem nodes used to create the edge weights allows for exploration of which texts create distinctive differences among the anthologies.

In co-citation analysis, an edge is created between two documents which are cited together in a third document (Small). In my co-printing network, which consists only of poem nodes, I create an edge between poems that are printed in the same anthology. Calculating modularity for this network reveals clusters of poems that frequently occur together in the same anthology. These clusters are made up of texts by multiple authors and can be used to explore components of canonicity, such as thematic, formal, or aesthetic qualities shared by poems in each cluster.

This paper argues that network analysis is a useful approach to examine the structure of the cultural field of Victorian poetry as it was constituted in key literary and teaching anthologies published from 1880-2002.

Bibliography

Bode, C. (2000). "Re-Definitions of the Canon of English Romantic Poetry in Recent Anthologies." *Anthologies of British Poetry: Critical Perspectives from Literary and Cultural Studies*. Ed. Barbara Korte, Ralf Schneider, and Stefanie Lethbridge. Amsterdam: Rodopi. 265-88.

Borgatti, S. (2009). 2-Mode Concepts in Social Network Analysis. *Encyclopedia of Complexity and System Science*. Medford: Springer, 8279-8291.

Bourdieu, P. (1993). "The Field of Cultural Production, or: The Economic World Reversed." In *The Field of Cultural Production: Essays on Art and Literature*. Edited by Randal Johnson. New York: Columbia UP. 29-73

Bourdieu, P. (1996) *The Rules of Art: Genesis and Structure of the Literary Field*. Translated by Susan Emanuel. Stanford: Stanford University Press.

Boyack, K. and Klavans, R. (2010). "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *Journal of the American Society for Information Science and Technology* 61.12 (2010): 2389-2404.

Damrosch, D. (2004). "From the Old World to the Whole World." *On Anthologies: Politics and Pedagogy*. Ed. Jeffrey Di Leo. Lincoln, NE: U of Nebraska P; 31-46.

Faust, K. (1997) "Centrality in Affiliation Networks." *Social Networks* 19: 157-191.

Fowler, A. (1979) "Genre and the Literary Canon." *New Literary History* 11.1 : 97-119.

Golding, A. C. (1984). "A History of American Poetry Anthologies." *Canons*. Ed. Robert von Hallberg. Chicago; London: U of Chicago P, 279-307.

Graff, G. (1992). *Beyond the Culture Wars: How Teaching the Conflicts can Revitalize American Education*. New York: W. W. Norton.

Guillory, J. (1993). *Cultural Capital: The Problem of Literary Canon Formation*. Chicago and London: The U of Chicago Press.

Harris, W. (1991). "Canonicity." *PMLA* 106: 110-21.

Kessler, M. M (1963). "Bibliographic coupling between scientific papers," *American Documentation* 24: 123-131.

Latané, D. (2000). "Recent Anthologies." Victorian Poetry 38.2: 331-339.

Lecker, R. (2013). *Keepers of the Code: English-Canadian Literary Anthologies and the Representation of the Nation*. Toronto: U of Toronto P, 2013.

Small, H. (1973). "Co-citation in the scientific literature: a new measure of the relationship between two documents" Journal of the American Society for Information Science 24: 265-269.

From Mnemosyne to Terpsichore – the Bilderatlas after the Image

Leonardo Impett
leonardo.impett@epfl.ch
École Polytechnique Fédérale de Lausanne, Switzerland

Sabine Susstrunk
sabine.susstrunk@epfl.ch
École Polytechnique Fédérale de Lausanne, Switzerland

Introduction

This study concerns Aby Warburg's last and most ambitious project: the Atlas Mnemosyne (or Bilderatlas), conceived in August 1926 and truncated three years later, unfinished, by Warburg's sudden death in October 1929. Mnemosyne consists of a series of large black panels, on which are attached a variable number of black-and-white photographs of paintings, sculptures, tarot cards, stamps, coins, and other types of images. The version we use is the one Warburg was working on at the time of his death, also known as the "1-79 version": it includes around a thousand images pinned on 63 panels (This version of the Atlas is published in various print editions, and available online). The Bilderatlas is a conceptual maze - the culmination of a life's scholarship in images and memory - through which perhaps the clearest thread is the concept of Pathosformel, or formula for (the expression of) passion. Much excellent work has been written on the concept but, to the best of our knowledge, nobody has yet tried to "operationalise" it - to turn it into a sequence of quantitative operations, or in other words, into an algorithm (Moretti, 2013).

The Pathosformel and its Operationalisation

On the most basic level, the Pathosformel describes the portrayal passionate emotions through a formula, a repeatable visual paradigm. The Pathosformel owes much of its force, as Salvatore Settis has pointed out, to its combination of semantic opposites: an "oxymoronic word, in that it merges in the same term the movement of pathos and the rigidity of the formula-schema" (Settis 1997).

Rather than attempting to operationalise the entire concept at once, we first break it down into its constituent components: first the morphology of formula, then the dynamism of pathos.

We turn to two well-studied cases of Warburgian formula: the Death of Orpheus (where Warburg first named the concept of Pathosformel), and the Nymph, headhuntress and Fortuna (Bilderatlas panels 46-48). Looking at Warburg's examples of the Orpheus-formula, reproduced in Figure 1, we can hardly stop ourselves from spotting a formula which repeats across the ages; but how could this intuitive similarity be measured?

Our algorithm comes as follows:

1. We isolate each individual body from its context. It is clear the Pathosformel relates to individual characters - the object of study thus becomes not panels or photographs in the Bilderatlas, but individual bodies.
2. We take only the skeletons of such bodies. Here we are eliminating colour, clothes, hands, faces, gender, age. This is not to say that such factors aren't important: but they are not elementary to the formula (see for example, André Jolles' letter to Warburg 23rd December 1900, where the formula of Ghirlandaio's Nymph hops between Judith, Salome, Tobias, Gabriel etc.– see Ghelardi, 2004; or the formula in Bilderatlas Panel 47 shared between Giambolo-gna's Samson and Donatello's Judith.)
3. We compare these skeletons by measuring the angles of the main limbs of the body, as described in Section 5.

Each of these steps is not merely a convenient quantification, but a conceptual wager. This is the strength of operationalisation as a critical tool: it forces one to be explicit about the conceptual choices one makes.

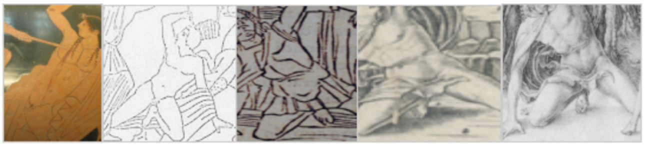


Fig 1: the Death of Orpheus: details from Warburg's example figures from classical antiquity to Dürer. Reproduced from Warburg 1905/1999

Anatomy, Emotion and Pose

Warburg was certainly influenced by Darwin's *The Expression of the Emotions in Man and Animals* (Darwin 1872) - which, when talking about human emotions, largely concentrates on the face. Indeed, the first figure of the book is an anatomical diagram of the face - "I shall often have to refer [...] to the muscles of the human face" (Darwin 1872 p.23). Warburg was certainly struck by the book - writing in his diary "ein Buch, das mir hilft!" (Gombrich, p.72). He was also interested and capable of studying the face in art (see e.g. his discussion of faces in Ghirlandaio's Confirmation in Santa Trinita, in The Art of Portraiture in the Florentine Bourgeoisie (in Warburg 1999 p.185), yet never in relation to Pathos - his descriptions of the Pathosformel relate exclusively to the *body*.

We can relate Warburg's decision to the large psychological literature on emotional recognition from bodies. Psychological studies are based on the Light Spots Model by Johansson (1973), often called 'biological motion', in which reducing body pose to 10-12 points - quite comparable to our own reduction - is judged to give a 'compelling impression of human walking, running, dancing etc.'. Using only Light Spots, observers can reliably tell gender and emotion from dynamic pose (Kozlowski 1977, Montepare 1987). Indeed, it has been suggested that our emotional understanding of faces is more influenced by our perception of the body than vice versa (Van den Stock 2007).

Encoding Pathos through Pose

The Atlas is, even by today's dataset standards, quite sizeable: 1000 images across 63 panels, containing an order of 10^3-10^4 depicted human figures. Scalable manual annotation is only therefore possible through crowdsourcing, which we did through the CrowdFlower platform.

Accurately annotating every visible figure in an image is a difficult and ambiguous tasks. Additionally, if different workers annotate different figures in the same image, the annotations cannot be collated or averaged. We therefore developed a two-stage annotation process:

Human figures are extracted from the painting by drawing bounding-boxes. This is done three times per image (by three separate workers).

Having aggregated the information from the first stage, separate images are produced for each human figure. Detailed pose information (the position of major body-points) is then added by three separate workers, with the information aggregated.

It should be clarified that the decision to annotate bodies in isolation (for annotation accuracy and just worker compensation) is quite separate to the earlier conceptual decision to analyse bodies individually, which relates to the object of study. It would be quite possible to do either one without the other.

Using this two-stage annotation process, we have presently annotated ⅓ of the Bilderatlas (by panels), resulting in 1,665 aggregated human poses. The collection and aggregation of the data are described in greater technical detail elsewhere (Impett, 2016).

Data Analysis: dimensionality reduction and dimensioned reductionism

Having encoded our static poses, how do we analyse and compare a collection of human figures of different sizes, proportions and orientations? We mirrored the poses horizontally and controlled for global rotation, ending up with a 11-dimensional vector *P*, describing the angles of the main limbs.

From this angular pose vector, we can use circular statistics to find a morphological distance $D_{a,b}$ between two poses P_a and P_b:

$$D_{a,\,b} = \sum_{i=1}^{10} |P_{a,\,i} - P_{b,\,i}|$$

Where $P_{a,\,i}$ is the ith angle of pose vector P_a , and $||_\alpha$ is the angular radian distance:

$$|x|_{\alpha} = min(x, \pi - x)$$

These morphological pose-differences are perceptually meaningful over short distances. On a larger scale, they become less perceptually significant: is a sitting person 'closer' to a lying or standing person?

In order to make our distance analysis perceptually relevant, therefore, we first clustered our 1,665 poses into 16 clusters by rotational K-means clustering (Dordet-Berdanet and Wicker 2008). Our two-stage clustering system is therefore as follows:

I. K-means clustering (to produce meaningful clusters)

II. Hierarchical clustering (for within-cluster morphological information)

The number of clusters K is chosen by looking at the inter-cluster variance over K. The result of the first stage of clustering is shown in Figure 2; Figure 3 shows an example section from a hierarchical map of Cluster 1.

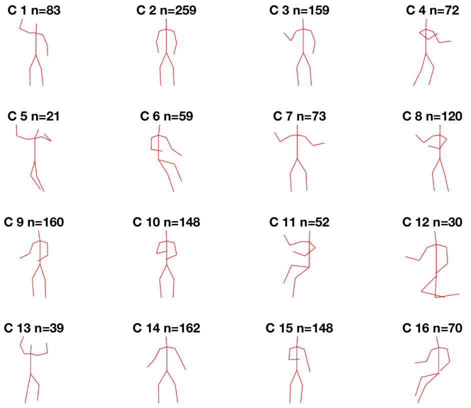


Fig 2: our 16 pose-clusters

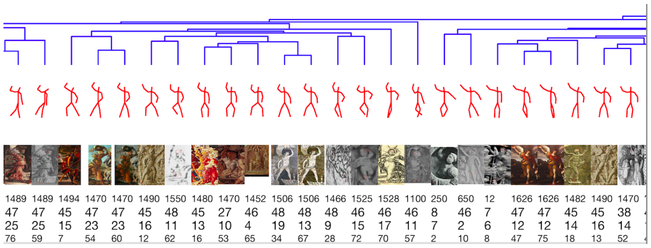


Fig 3: a detail from a dendrogram of different poses within the Bilderatlas, produced by second-stage clustering within Cluster 2

Unity of the Pathosformeln: from distant to close reading

Some of the clusters clearly represent physical activities - sitting, praying, embracing, dancing - whilst others seem more subtly communicative or expressive in nature. Having reorganised the 1,665 figures into 16 mean-centred clusters, we proceeded to trace the classical Pathosformeln - identified in the Atlas by primary and secondary literature - through our clusters.

The canonical Pathosformeln are mainly mythological figures (Perseus, Pentheus, Orpheus) or recurring allegories (Graces, Nymphs, Fortuna). They were previously described as distinct, and we expected to find a taxonomy of such formulae through our clustering analysis.

On the contrary, the statistical result was much stronger: a complete morphological unity. Almost every identified Pathosformel falls into Cluster 1, with few false positives - over 80% of the figures in Cluster 1 are an identified Pathosformel. The handful of exceptions are all borderline cases, placed in peripheral to Cluster 1 (Clusters 7 and 13).

Looking more closely at the images themselves, as in Figure 4, this becomes visually clear: not only do the Pathosformeln share certain pose features (most importantly, a raised arm) present nowhere else in the dataset. To date, however, the authors know of no art-historical literature that has identified such morphological unity.

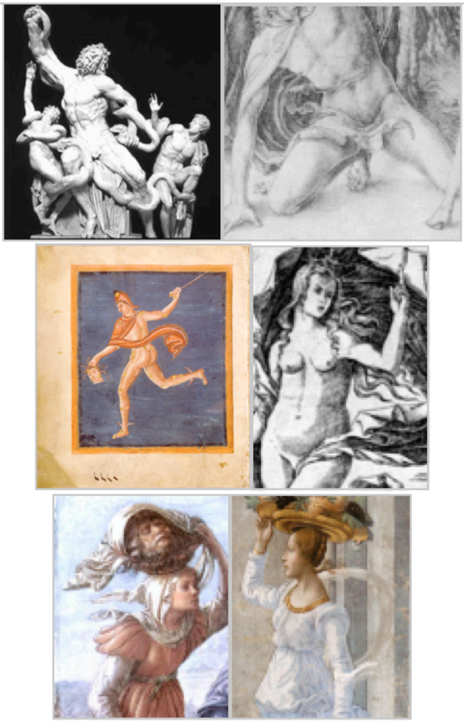


Figure 4, clockwise from top-left: Laocoön, Orpheus, Fortuna, Nymph, Judith, Perseus - all except Orpheus are in the Atlas. The identified Pathosformeln share distinguishing features from the other characters in the Atlas: a raised arm, most often accompanied by a lowered second arm, and a slight twist of the body.

Concluding remarks

Our morphological model for Pathosformel is statistically strong: but what are the art-historical implications? The oppositional symmetry and raised arm of Cluster 1 (Fig. 2) reminds us of a Contrapposto, but the bodies themselves are far removed from such classical balance (e.g. Fig. 4, top). Rather than movement, tension (between upper and lower body) seems to be the fundamental element of Pathosformel - the nature of which will be the subject of a subsequent publication.

Our morphological analysis has shown that static pose can identify Pathosformeln, and that a study of static pose through a large collection of artistic works can identify links across styles, periods and cultures.

Automatic detection of pose is a focus of the current research, and will allow us to expand our art-historical models beyond the Bilderatlas. Humans (thus bodies, and poses) are unsurprisingly the most common feature of human art, and therefore make excellent objects of study for an art history of the Longue Durée (Robb 2015).

Bibliography

Darwin, C. (1872). The expression of the emotions in man and animals. London, John Murray

Dortet-Bernadet, J.L. and Wicker, N., (2008). Model-based clustering on the unit sphere with an illustration using gene expression profiles. In Biostatistics, 9(1), pp.66-80.

Gombrich, E.H. (1970). Aby Warburg, an intellectual biography. Chicago University Press.

Impett, L. and Süsstrunk, S., (2016). Pose and Pathosformel in Aby Warburg's Bilderatlas. In European Conference on Computer Vision (pp. 888-902). Springer International Publishing.

Johansson, G., (1973). Visual perception of biological motion and a model for its analysis. In Perception & psychophysics, 14(2), pp.201-211.

Kozlowski, L.T. and Cutting, J.E., (1977). Recognizing the sex of a walker from a dynamic point-light display. In Perception & Psychophysics, 21(6), pp.575-580.

Montepare, J.M., Goldstein, S.B. and Clausen, A., (1987). The identification of emotions from gait information. In Journal of Nonverbal Behavior, 11(1), pp.33-42.

Moretti, F., (2013). "Operationalizing"; or, The Function of Measurement in Modern Literary. Stanford Literary Lab Pamphlet Series. litlab.stanford. edu/LiteraryLabPamphlet6

Robb, J. and Harris, O.J.T. (2013). "Body worlds and their history: some working concepts." The Body in History: Europe from the Palaeolithic to the future: 7(31).

Settis, "Pathos und Ethos, Morphologie und Funktion", in W. Kemp, G. Mattenklott, M. Wagner, M. Warnke, eds, Vorträge aus dem Warburg-Haus, Band 1, De Gruyter, Berlin 1997, pp. 39-44.

Van den Stock, J., Righart, R. and De Gelder, B., (2007). Body expressions influence recognition of emotions in the face and voice. In Emotion, 7(3), p.487.

Warburg, A., trans. D. Britt. (1999) The Renewal of Pagan Antiquity. Los Angeles, Getty Research Institute.

Interactive Visual Exploration of the Regesta Imperii

Markus John
markus.john@vis.uni-stuttgart.de
University of Stuttgart, Germany

Christian Richter
christian.richter.rr@gmail.com
University of Stuttgart, Germany

Steffen Koch
steffen.koch@vis.uni-stuttgart.de
University of Stuttgart, Germany

Andreas Kuczera
kuczera@uni-mainz.de
University of Mainz, Germany

Thomas Ertl
thomas.ertl@vis.uni-stuttgart.de
University of Stuttgart, Germany

Introduction

The Academy of Sciences and Literature in Mainz provides online access to the Regesta Imperii - a very extensive historical dataset based on documentary sources of German-Roman kings. About 125,000 regestae of emperors and popes are searchable and viewable in this online portal. The current user interface offers direct access to single documents through different form-based search facilities, as well as through a catalogue that directly reflects the structure of the regestae volumes as they have been created in this long-term project. In order to further improve access to this large data volume, we suggest an additional approach based on coordinated views. The usage of coordinated view approaches is very common in many domains (Stasko et al, 2008, Vuillemot et al, 2009, Koch et al, 2011). However, there is no publicly system available, which would provide a suitable access to this historical dataset. The motivation for this new approach is twofold. On the one hand, we improve the support for search and exploration tasks in this historical data set that are based on imprecise information needs or on a less deep understanding of the available information. In practice, such imprecise queries can quickly lead to an overwhelmingly large number of search results. Allowing users to create and refine queries in a visual way, while offering immediate feedback on the number of entries requested, can help to cope with underspecified queries and help refining them iteratively (Jänicke et al, 2012). On the other hand, we offer a powerful means for visually analyzing the available information and understanding complex relationships by providing different linked perspectives on subsets of the collections. These perspectives include views on historic persons and entities as well as temporal and spatial information contained in the regestae. A usage scenario shows successful application of the approach.

Visual approach

We implemented a web-based visualization that is easily available to humanities scholars, since it does not require users to install software. The web-based visualization uses the library Data-Driven Documents (D3) (Bostock et al, 2011) and runs with a web browser supporting HTML5, SVG, CSS, and JavaScript.

Data processing

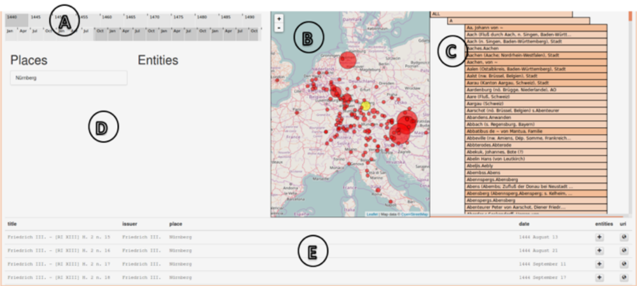
In the Regesta Imperii database all documentary sources of German-Roman kings are available in full text (xml-files). This data set consists of regestae volumes and register information. Regestae volumes are short summaries of a text and contain important metadata of a document, such as the title of the document, an ID for the unique assignment, date of issue, and place of issue as name and coordinates.

Since places, persons, and institutions may be known by different names, entries can be overlooked within the full-text search. Therefore, place and personal registers provide a central resource for work with the regestae and contain a list of places, persons, institutions, and additional information, such as the numbers of the regestae in which the entity is mentioned, a reference to another entity entry (if available), and a unique id.

Since the regestae have been manually digitized, the xml-files can contain syntax or other transmission errors,such as different date formats, geo-coordinates or tags. Therefore, the data must be parsed and well prepared in order to use them for a visual analysis.

Visual approach

After the regestae and register volumes have been successfully parsed and loaded into our system, users can start their exploration in the main view as depicted in Figure 1. The main view consists of the five coordinated views: (A) timeline view, (B) map view, (C) register view, (D) overview filter view, and (E) results view.



Aa, Johann von –
Aach (Fluß durch Aach, n. Singen, Baden-Württ...
Aach (n. Singen, Baden-Württemberg), Stadt
Aaches.Aachen
Aachen (Aache; Nordrhein-Westfalen), Stadt
Aachen, von –
Aalen (Ostalbkreis, Baden-Württemberg), Stadt
Aalst (nw. Brüssel, Belgien), Stadt
Aarau (Kanton Aargau, Schweiz), Stadt
Aardenburg (nö. Brügge, Niederlande), AO
Aare (Fluß, Schweiz)
Aargau (Schweiz)
Aarschot (nö. Brüssel, Belgien) s.Abenteurer
Abandens.Anwanden
Abbach (s. Regensburg, Bayern)
Abbatibus de – von Mantua, Familie
Abbeville (nw. Amiens, Dép. Somme, Frankreich...
Abbterodes.Abterode
Abekuk, Johannes, Bote (?)
Abelin Hans (von Leutkirch)
Abeljis.Aebly
Abembss.Abens
Abennspergs.Abensberg
Abens (Abembs; Zufluß der Donau bei Neustadt ...
Abensberg (Abennsperg,Abensperg; s. Kelheim, ...
Abenspergs.Abensberg
Abenteurer Peter von Aarschot, Diener Friedr...

title	issuer	place	date	entities	uri
Friedrich III. - [RI XIII] H. 2 n. 15	Friedrich III.	Nürnberg	1444 August 13	+	●
Friedrich III. - [RI XIII] H. 2 n. 16	Friedrich III.	Nürnberg	1444 August 21	+	●
Friedrich III. - [RI XIII] H. 2 n. 17	Friedrich III.	Nürnberg	1444 September 11	+	●
Friedrich III. - [RI XIII] H. 2 n. 18	Friedrich III.	Nürnberg	1444 September 17	+	●

Figure 1. The interactive visualization approach for exploring and analyzing regestae of royal and papal records.

We initially depict all available data and enable users to create and refine queries in a visual way iteratively, to reduce the overall set. For example, users can select certain time periods and places in which the regestae were published or persons and places that are mentioned in the regestae.

The timeline view consists of two stacked timelines and enables users to select a time period by clicking and dragging, as depicted in Figure 6, comparable to the Simile approach (Huynh, 2008). The first timeline allows a coarse filtering and represents the whole-time range of the regestae. Once a timespan is selected, the second timeline is updated and permits a finer selection of the upper selected time range.

To get an overview where the regestae were published, users can discover the map view as depicted in Figure 2. The map view uses the JavaScript library Leaflet (Agafonkin, 2014) which supports interactive features such as zooming and panning. The red circles in the map represent places where the regestae were published and the circles size is scaled proportionally to the places occurrences, similar to the approach (DARIAH-DE, 2015). This helps to get a quick overview of important places. When hovering over a circle, a tooltip shows the corresponding place name. By selecting one or more circles (highlighted in yellow), the places are added to the search query.

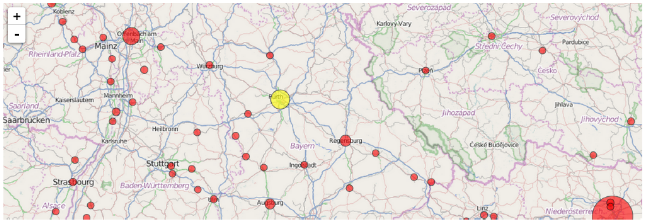
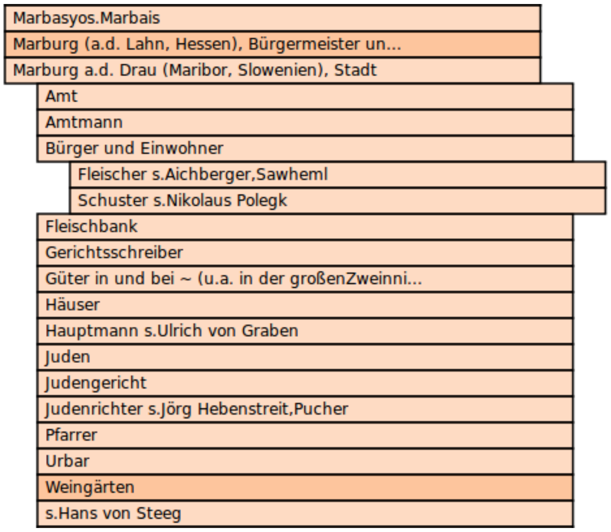


Figure 2. The map view gives an overview where the regestae were published.

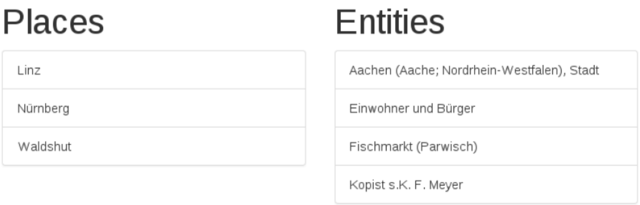
The register view represents the entities from the register volumes in an alphabetically sorted hierarchical structure as depicted in Figure 3. Users can explore the different entries by clicking on the different hierarchies. Nodes, which contain further entities are displayed in darker color. Furthermore, users can select one or more entities to adapt the search query.



Marbasyos.Marbais
Marburg (a.d. Lahn, Hessen), Bürgermeister un...
Marburg a.d. Drau (Maribor, Slowenien), Stadt
Amt
Amtmann
Bürger und Einwohner
Fleischer s.Aichberger,Sawheml
Schuster s.Nikolaus Polegk
Fleischbank
Gerichtsschreiber
Güter in und bei ~ (u.a. in der großenZweinni...
Häuser
Hauptmann s.Ulrich von Graben
Juden
Judengericht
Judenrichter s.Jörg Hebenstreit,Pucher
Pfarrer
Urbar
Weingärten
s.Hans von Steeg

Figure 3. The register view represents all persons, places, and institutions which are mentioned in the regestae.

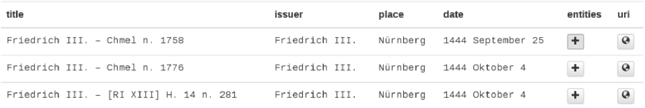
From the overview filter view, users can get a summary about all the selected places and entities which determine the search results. In addition, users can deselect places and entities from the list to adapt the search query.



Places	Entities
Linz	Aachen (Aache; Nordrhein-Westfalen), Stadt
Nürnberg	Einwohner und Bürger
Waldshut	Fischmarkt (Parwisch)
	Kopist s.K. F. Meyer

Figure 4. The overview filter view gives an overview of all selected places and entities.

Based on the combined search query, the result view lists all regestae entries, which are included in the search results as depicted in Figure 5. For each entry, the list displays the following metadata: title of the regesta, issuer, place and date of issue. By clicking on the icon in the column entities, users get the information of all entities occurring in the regesta in a separate list view. Furthermore, users can select the icon in the last column uri to jump directly to the corresponding regesta entry on the web page of the Regesta Imperii. This enables users to analyze the regesta in detail.



title	issuer	place	date	entities	uri
Friedrich III. - Chmel n. 1758	Friedrich III.	Nürnberg	1444 September 25	+	●
Friedrich III. - Chmel n. 1776	Friedrich III.	Nürnberg	1444 Oktober 4	+	●
Friedrich III. - [RI XIII] H. 14 n. 281	Friedrich III.	Nürnberg	1444 Oktober 4	+	●

Figure 5. The result view displays all regestae entries which are included in the result set.

Usage scenario

In the following section, we present a usage scenario that occurred during one of the joint workshops of users from the Academy of Sciences and Literature in Mainz and the developers of the approach. Therefore, our usage scenario represents instead the insights and lessons learned through several sessions with two experts from the Regesta Imperii.

In a first step, the expert explores and analyze the map view. That way, she gets a quick overview of the places where regestae were published. During the exploration, the place Nürnberg (Nuremberg) has aroused her interest, since she has already examined these entries a long time ago. To find out more about the regestae entries, she starts a search query by selecting Nürnberg (Figure 6A) in the map view and discovers the search results in result view. However, the result set is too large for a further deeper analysis. Therefore, she refines the search query by selecting the time period from 1440 to 1450 in the timeline view (Figure 6B), because she is especially interested in the early regestae entries. As the next step, she searches for entities that are mentioned in the regestae with the aid of the results view. She finds out that there are many connections to French entities. To further analyze that, she selects the hierarchy “Frankreich, Königreich” (France, Kingdom) in the register view, as depicted in Figure 6C. By adjusting the search query, she received a specified subset of the collection for a further analysis. This way, she finds that primarily French kings are mentioned in the regestae entries.

During the analysis, she learns from the map view that Neustadt near Bremen (Figure 6D) has many regestae entries which she did not expect. To inspect this in detail, she selects Neustadt and explores the result list. By analyzing the different entries in the list and web page of the Regesta Imperii, she finds out that the geo-coordinates were manually digitized incorrectly. Consequently, the expert corrects the entries in the database by assigning these entries to the actual place Neustadt near Vienna.

While these sessions, we received a lot of feedback that showed that our approach improves access to the large data volumes of the Regesta Imperii and facilitates search and exploration tasks, as well as assisting in understanding complex relationships and gaining new insights.

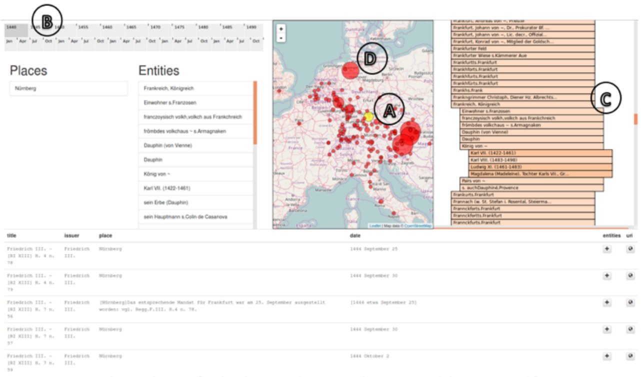


Figure 6. An example search query for the place Nürnberg, French entities, and the time period from 1400 to 1440.

Conclusion and future work

The presented interactive web-based approach has been evaluated through expert feedback that recommends it as an effective method for exploration analysis.

We are planning to extend the different linked views to support users with additional information. Concerning this issue, we have implemented the relative distribution of the regestae volumes in the timeline view, as depicted in Figure 6, and we are currently working on the co-highlighting between the views. We will also ensure that experts from the Regesta Imperii are able to correct errors that arise during the digitization process interactively.

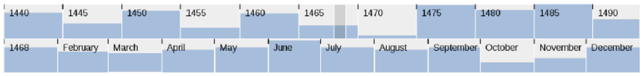


Figure 7. Timeline filter view of the selected year 1468 with a relative distribution of the regestae over time.

Bibliography

Stasko, J., Görg, C., and Liu. Z. (2008). Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization,* 7(2):118– 132.

Vuillemot, R., Clement, T., Plaisant, C., and Kumar, A. (2009) What's being said near "martha"? exploring name entities in literary text collections. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology,* 2009. VAST 2009., pages 107–114.

Koch, S., Bosch, Giereth, H. M., and T. Ertl, T. (2011). "Iterative Integration of Visual Insights during Scalable Patent Search and Analysis," in *IEEE Transactions on Visualization and Computer Graphics,* vol. 17, no. 5, pp. 557-569.

Jänicke, S., Heine, C., Stockmann, R., and Scheuermann, G. (2012). Comparative visualization of geospatialtemporal data. In *Proceedings of the 3rd International Conference on Information Visualization Theory and Applications,* IVAPP, pages 613–625.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). "D^3 data-driven documents." *IEEE transactions on visualization and computer graphics:* 2301-2309.

Huynh, D. (2008). SIMILE—Timeline. Simile Projects, http://www. simile-widgets. org/timeline/ (accessed 28.10.2016).

Agafonkin, V. (2014). "Leaflet. an open-source javascript library for mobile-friendly interactive maps." (accessed 28.10.2016).

DARIAH-DE. (2015). Geo-Browser: https://geobrowser.de.dariah.eu/(accessed 28.10.2016).

Reverse Engineering the First Humanities Computing Center

Steven E. Jones
stephenjones@usf.edu
University of South Florida, United States of America

Julianne Nyhan
j.nyhan@ucl.ac.uk
University College London, England

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
University of Alberta, Canada

Stéfan Sinclair
stefan.sinclair@mcgill.ca
McGill University, Canada

Melissa Terras
m.terras@ucl.ac.uk
University College London, England

How can digital methods be used to conceptualize historical research projects, including their teams, approaches, methods, and outputs? What methodologies can be used to synthesize and analyze archival records, workshop plans, photographic evidence, and oral histories? This co-authored paper describes an ongoing effort by a collaborative group* to understand and recover the work of Father Roberto Busa, commonly thought to be the "founding father" of Digital Humanities. Starting in 1949, Roberto Busa, S.J., began a landmark collaboration with IBM to build a lemmatized concordance to the works of St. Thomas Aquinas. In 1956 Busa founded the world's first humanities computing center in Gallarate, Italy, located after 1961 in a former textile factory stocked with rows of IBM punched-card machines. This was CAAL, the *Centro per L'Automazione dell'Analisi Letteraria*—the Center for the Automation of Literary Analysis. There Busa and his mostly female student operators processed the monumental *Index Thomisticus*, a selection of the Dead Sea Scrolls, and other texts, from 1961-1967 (Terras and Nyhan, 2016; Jones 2016). However, there is much that is still unknown about Busa's research approach and methods. We aim to recover what Busa and his operators did. By repurposing punched-card office machinery for *literary* data processing, Busa created an important pre-computing technology platform for humanities research, one which has become obscured over time. We aim to reverse engineer and reconstruct not just a particular technology (punched-card machines) but that first humanities computing center as a whole. By doing so, we will explore methods that can be useful for other historians as they look back upon site-specific projects and groups, using digital tools and methods to effectively interleave and investigate historical data sources.

Jeffrey Schnapp of Harvard's metaLAB has remarked that "every cultural object is a network" (Schnapp, 2015). Reverse engineering involves taking apart a device or system not to replicate it but in order to better understand its design and purpose, its networked relations. The goal in this case is to break down and experimentally reconstruct the networked cultural objects--including specific machines, architecture, infrastructure, and human operators--that amount to the components of that Italian humanities computing center, and in that way to model a more capacious idea of "digitization" itself. We make use of a cluster of convergent practices:

1. The digitization of archives--paper-based documents with Dublin Core derived metadata, but

also 3D digitizations of physical artifacts such as punched cards, relay switches, etc.

2. A cultural-heritage virtual model of the architectural space, a 3D immersive environments of the center itself, created through basic photogrammetry and using Maya + the Unity engine, based on multiple archival images, as well as new scans of the building, still standing outside Milan (though much altered).
3. Emulations of forgotten or obsolete technologies, punched-card data processing systems as well as other "adjacent" technologies in the 1950s and 1960s.
4. Oral histories and audio-files of interviews with surviving punched-card operators, Busa's secretaries, and others

The overall objective is to model this important early research center and its activities through a series of purpose-driven and interlinked emulations, 3D spaces, oral histories, and digitized documents and artifacts. We employ metadata to map archival materials and emulations onto the models in order to understand the material history of what is usually taken to be the first humanities computing center. In the process, we complicate the key terms themselves, including *first* and *computing*.)

Most of what has been known to date about Busa's early literary data processing was derived from a handful of his own publications– first by Winter (1999); later, Jones (2016) drew on the Busa Archive to contextualize and extend his narrative account. Rockwell and Sinclair (2014) and Terras and Nyhan (2016), have continued to clarify the story in different ways. Actually modeling the machinery and workflow allows us to address specific questions about this important moment in the birth of linguistic data processing and humanities computing, such as:

- What were the precise roles played by human operators *between* the automated stages, sorting card decks, lemmatizing word lists, programming machines via plugboards, etc.? (How were these roles stratified and gendered?)
- What source texts were used for input and how were they prepared and marked up so that the operators could use them as the basis for what they punched on the cards?
- At what stage did IBM agree to print customized punched cards with what amounted to data fields unique to Busa's projects? What was the nature of the data ontology behind these customizations?
- What is the evidence that the work of Busa's center contributed to larger technology developments at IBM, such as Peter Luhn's development of the influential KWIC (keyword in context) protocol for information retrieval?

Additional questions will surely arise during the ongoing process of modeling and cross-checking archival materials and oral histories.

Although Busa's humanities computing center is our focus, we believe this methodological approach would be useful in other instances, as a way to conceive of digitization as a process of modeling artifacts and documents in relation to technology and infrastructure. We draw on theoretical approaches and methods associated with media archaeology (Parikka, 2012; Emerson, 2014; Rockwell and Sinclair, 2014; Sinclair, 2016), creative historical prototyping (U Victoria Maker Lab; Sayers et al, 2016), the archaeology of science (Haigh, 2016; Schiffer, 2001), and on the methods and expertise of digital archaeology in the field of cultural heritage, including its attention to issues of access and preservation (Koller, 2009; London Charter, 2009).

The presentation at DH 2017 will include slides containing selections from the 800 historical photographs of Busa's center, as well as other images, audio files, and demonstrations, including a prototype 3D virtual model of the center. The paper will explain the project's practical aims and theoretical significance: for example, we address current debates in digital humanities about the influence of text-based analysis on today's definitions and practices; or debates about possible alternative genealogies for DH (Klein, 2012). It will also spotlight the role of gendered labor in early humanities computing, and the entanglements of early humanities technology research with corporate and government funding. Our broader methodological purpose is to take up in practice what Jeffrey Schnapp has called the "defining design challenge of our epoch"—"to weave together information and space in a meaningful fashion" (Schnapp, 2015), and the methods will be of interest and use to others who are approaching multimodal archives and interpolating the information therein.

Acknowledgements

In addition to the coauthors of this paper, contributors to the overall project include Marco Passarotti and Paolo Senna (*Università Cattolica del Sacro Cuore*, Milan).

Bibliography

Emerson, L. (2014). *Reading Writing Interfaces: From the Digital to the Bookboun*d. Minneapolis: University of Minnesota Press

Haigh, T., Priestley, M., and Rope, C. (2016). *Eniac In Action: Making & Remaking the Modern Computer*. Cambridge, MA: MIT Press.

Jones, S.E. (2016). *Roberto Busa, S.J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York: Routledge.

Klein, L. (2012). "Digital Origin Stories." http://lklein.com/2012/11/digital-origin-stories/.

Koller, D., Frischer, B., and Humphreys, G. (2009). "Research challenges for digital archives of 3D cultural heritage models." *ACM Journal* (December 2009): DOI: 10.1145/1658346.1658347 http://doi.acm.org/10.1145/1658346.1658347.

London Charter for the computer-based visualization of cultural heritage, 2.1 (February 2009): http://www.londoncharter.org.

Parikka, J. (2012). *What is Media Archaeology?* Cambridge, UK: Polity.

Rockwell, G., and Sinclair, S. (2014). "Past Analytical: Towards an Archaeology of Text Analysis Tools." *Digital Humanities 2014: Conference Abstracts*. Lausanne: EPFL and UNIL, pp. 359-60.

http://www.researchgate.net/publication/273449857_Towards_an_ Archaeology_of_Text_Analysis_Tools.

Sayers, J., Elliott, D., Kraus, K., Nowviskie, B., and Turkel, W. J. (2016). "Between Bits and Atoms: Physical Computing and Desktop Fabrication in the Humanities." In Schreibman, S., Siemens, R., and Unsworth, J., Eds. *A New Companion to Digital Humanities,* 3-21. Wiley Blackwell.

Schiffer, M. B., Ed. (2001). *Anthropological Perspectives on Technology*. Salt Lake City: University of Utah Press.

Schnapp, J. (2015). "Aphorisms on the 21st Century Museum." http://jeffreyschnapp.com/2015/01/26/aphorisms-on-the-21st-century-museum/.

Sinclair, S. (2016) "Experiments with Punch Cards." http://stefansinclair.name/punchcard/.

Terras, M. and Nyhan, J. (2016). "Father Busa's Female Punched-Card Operators." In Gold, Matthew K., and Klein, L. Eds. *Debates in Digital Humanities 2016.* Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/57.

University of Victoria Maker Lab. http://maker.uvic.ca.

Winter, T. N. (1999). "Roberto Busa, S.J., and the Invention of the Machine-Generated Concordance." *The Classical Bulletin,* 75,1: 3–20.

Personality and Politics: Myers–Briggs Personality Types on Twitter in the US 2016 Presidential Election

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, United States of America

Sean Vinsick
sean.vinsick@gmail.com
Duquesne University, United States of America

As a social media platform, Twitter (twitter.com) offers opportunities to examine public reactions in what could be described as the world's largest village square. Twitter provides an opportunity not only to participate in public discourse, but even to shape it. However, to fully understand this discourse, it is helpful to understand the people behind the discourse. Stylometric authorship profiling (Argamon, et al, 2009) provides an example of this. People are free to post their opinions, but by analyzing the writing style of the individual tweets, we can infer other attributes of the individual authors.

In this study, we infer personality categories using the well-known Myers-Briggs Type Inventory (MBTI) to analyze whether or not there are any differences between the supporters (and detractors) of the major party candidates, Hillary Clinton (D) and Donald Trump (R). The MBTI categorizes people along four major axes, as encoded in a four-character summary (for example, ENFJ: Extroverted, iNtuitive, Feeling, Judging). We have shown (Gray and Juola, 2011; Juola et al., 2013) that personality can be inferred with high accuracy from writing, and further that MBTI personality types can be gleaned specifically from Twitter feeds. Using The EthosIO system developed by Juola & Associates, we have applied this (Juola, Vinsick, and Ryan, 2016) to large-scale analyses of the demographics of personality on Twitter, finding substantial differences between the accepted distribution of personality in the general US population and between the distribution of personality among active Twitter participants. For example, introverts make up approximately 50% of the general US population, but nearly 80% of active Twitter users. Similarly, nearly 3/5 of the general US population are "sensing" (S) [as opposed to "intuitive" (N)], but half or fewer Twitter users are. Two specific subgroups, INFP and INFJ, are vastly overrepresented on Twitter, being only about 5% of the overall US population, but 30% or more of the samples gleaned from Twitter.

We extend this to analyzing personality differences between politically disparate groups of Twitter participants. As with (Juola, Vinsick, and Ryan, 2016), we harvested a large group of user names from the Twitter sample feed, selecting users whose public tweets included one of several political hashtags. Based on the hashtags seen, we divided participants into four groups: anti-Clinton (identified by one or more of '#NeverHillary', '#CrookedHillary', '#WhichHillary', '#DraftOurDaughters', '#hillary4prison', '#hillary4prison2016', '#StopHillary', '#CrimeWithHer', or '#Killary'); pro-Clinton ('#ImWithHer', '#Clinton', '#ClintonKaine16', '#ClintonKaine2016', '#HillaryClinton'], '#ClintonKaine', '#WhyIWantHillary', '#HillarysArmy', or '#Hillary2016'); anti-Trump ('#NeverTrump', '#dumptrump', '#trumptaxreturns', '#dontvotefortrump', '#dumpthetrump', '#boycotttrump', '#trumpsexism'); and pro-Trump ('#ImWithYou', '#TrumpTrain', '#MakeAmericaGreatAgain', '#TrumpPence16', '#TrumpPence2016', '#Trump', '#AltRight', '#VoteTrump', '#TeamTrump') This gave us approximately 600 user names for each of the four groups in our preliminary dataset. These user names were submitted to the EthosIO personality analyzer to produce distributional data for each subgroups.

We therefore had in our preliminary corpus 651 pro-Clinton subjects, 587 pro-Trump subjects, 635 anti-Clinton subjects and 639 anti-Trump subjects, for a total of 2512 total user names, divided across 16 MBTI categories (details in full paper). As expected from previous work, the overall statistics do not match US demographics; for example, types ISFP and INFP are strongly overrepresented in all samples, as are introverts in general. Our interest, however, is in whether or not political differences also show up as personality differences as well. In plainer language, does the average Clinton supporter have a different personality than the average Trump supporter?

We tested this hypothesis with a variety of chi-squared tests (df=15 throughout). At the most basic level, we found extremely significant differences (p ~ 10^-9) between Clinton supporters and Trump supporters. We also found significant differences (p ~ 10^-12) between "Democrats" (either pro-Clinton or anti-Trump) and "Republicans" (pro-Trump and anti-Clinton). Examining cells in detail suggest that ISFJ and ISFP are both overrepresented among Democrats while ESTJs and INFJs are overrepresented among Republicans.

By contrast, there was no significant difference (p > 0.10) between "Anti" and "Pro" subjects, despite the possible participation, for example, of third party supporters who are opposed to both Trump and Clinton. Similarly, we found no significant difference between pro-Trump and anti-Clinton subjects, or between anti-Trump and pro-Clinton subjects, suggesting that other factors than personality are affecting whether a person chooses to self-express in favor of a particular candidate or in opposition to that candidate's rival(s).

In this study, there are a number of potential confounding factors, the effects of which have not yet been assessed. The first is simply the presence of active attempts to manipulate the dialogue, for example, through the use of automated 'bots' (Kollanyi, Howard, and Wolley, 2016), or simply through the use of "sock puppets," multiple identities in an attempt to create a an appearance of consensus and of larger margins. A second factor is the issue of overlapping categories. Approximately 20% of our preliminary "anti-Trump" sample also self-identified as "pro-Clinton," and similarly, approximately 20% of the anti-Clinton sample self-identified as pro-Trump. More counterintuitively, approximately 5% of the anti-Clinton sample also identified as anti-Trump, and approximately 5% of the pro-Clinton sample was also pro-Trump. This may be due to a third confounding factor, the inability of simple keyword spotting to identify the use of irony (for example, in posting a link to an article highly critical of Trump's campaign and using the '#Trump' hashtag to draw attention to it). As we continue this analysis, based in part on data to be collected during the final and most intense week of the campaign, we will address these issues (and discuss our methods of address in the final paper).

We have therefore shown, using text analysis of Twitter on a moderately large scale, that there are significant differences between the types of people who self-identify as supporters (and opponents) of one of the major candidates in the 2016 US presidential election. We have also shown that there does not appear to be significant personality-related differences between whether one supports one's chosen candidate or opposes the other one. We have also confirmed the previous results (Juola, Vinsick, and Ryan, 2016) about the general distribution of personality types on Twitter, and hasten to point out that the differences we have identified are still relatively minor and that the overall distribution of personality types in both camps are broadly similar to the distribution of personality types on Twitter in general. However, our results show, first, that, in keeping with prior work, inferring personality type via Twitter is practical and useful. Second, they show that personality may play a factor in the selection of one's chosen candidate.

Finally, the question of "who are Trump's voters?" against "who are Clinton's voters?" will no doubt interest historians for decades. Our results provide some insight into possible psychological motivations in addition to the more traditional social, political, and economics reasons, and may therefore enrich future discussion and scholarship.

Nota bene

This version of the paper was written approximately one week before the actual 2016 election and will be updated as appropriate.

Network analysis of the manuscript context of Old Icelandic literature

Katarzyna Anna Kapitan
kak@hum.ku.dk
University of Copenhagen, Denmark

Aim

This paper explores the possibilities of applying computer-assisted methods to the field of Nordic manuscript studies, with a special emphasis on a network analysis—in a broad sense—of manuscript context. A case study of one Icelandic legendary saga's manuscript tradition is used to test the hypothesis that the manuscript context can carry information about ethnic genre associations of the text (on ethnic genre in Old Norse literature see: Harris, 1975; on legendary sagas as a genre see: Quinn, 2006).

Research Questions

Hrómundar saga Gripssonar traditionally belongs in the corpus of legendary sagas (fornaldarsögur); it was included in the second volume of Rafn's (1829) *Fornaldarsögur Norðrlanda*, and in Björner's (1737) *Nordiska kämpa dater.* The saga as it is known today, however, is a post-medieval re-working of a metrical version of the story known as rímur (Brown, 1946), and is probably not much older than seventeenth century. Therefore it does not necessarily fit well with the other texts included the corpus of legendary sagas, as they usually date from the fourteenth and fifteenth centuries (Driscoll, 2005:207). This makes *Hrómundar saga Gripssonar* an interesting case study for investigations of the text's genre affiliation in the extant manuscripts preserving the saga. Does it appear frequently in manuscripts with the older legendary sagas or with younger rímur-based narratives? To answer this question, I first examine the manuscript context of legendary sagas as a corpus, based on collaborative research with Rowbotham and Wills (Kapitan et al., 2017); second, I examine the position of *Hrómundar saga Gripssonar* within the corpus and its relationships with other texts.

State of the Art

Much discussion in the field of Old Norse studies centers on whether the legendary sagas deserve to be considered a separate literary genre, or should instead be analyzed as chivalric literature (Quinn, 2006). One of the main reasons for these considerations seems to be the fact that the term fornaldarsögur is not attested in the medieval texts; it was introduced in the early nineteenth century by C.C. Rafn, who published a collection of texts under the title *Fornaldarsögur Norðrlanda* (Rafn, 1829). Rafn's selection of texts and the definition of fornaldarsögur as a corpus of texts dealing with events taking place in Scandinavia before the settlement of Iceland, however, was not detached from pervious scholarship of early eighteenth century (Lavender, 2015). The current discussion on the legendary sagas as a corpus (or a genre) is polarized around contradicting opinions. Some scholars suggest that the legendary sagas had to be considered a separate category in pre-modern period, because they are frequently bound together in the manuscripts (Guðmundsdóttir, 2001:cxlvii; Mitchell, 1991:21) while others, using the same argument, emphasize strong connections between the legendary sagas and the chivalric sagas (Driscoll, 2005:193). Additional problems arise when classifying the generic hybrids (Rowe 1993; 2004) appearing within the corpus, or distinguishing between the legendary sagas (fornaldarsögur) and the late legendary sagas (fornaldarsögur síðari tíma; Driscoll 2005). Even though scholars eagerly turn towards the manuscript context to support their claims regarding the genre classification, no comprehensive overview of the legendary sagas' codicological context has yet been presented. This gave rise to the project *Stories for all times* conducted at the University of Copenhagen, which created a complete catalogue of manuscripts preserving legendary sagas. The catalogue contains 818 TEI-conformant XML-based manuscript descriptions with over 8000 items, 1764 of which are classified as legendary sagas and 920 as chivalric sagas. This amount of data is much too large to be analyzed manually, therefore it is necessary to apply computer-assisted analysis in order to draw some general conclusions regarding this corpus and the relationships between these texts.

Methods

The first part of my paper, which aims to establish the position of *Hrómundar saga Gripssonar* within the wider context of the manuscript, draws on the network analysis of the corpus, conducted in collaboration with Rowbotham and Wills (Kapitan et al., 2017). There, the codicological context of a text was considered as a system of relationships between texts, and following Hall's (2013) approach in his network of chivalric sagas, texts were represented as nodes, manuscripts as edges, both visualized with the free visualization software Gephi. The second part is based on database queries aimed at obtaining detailed information about particular manuscripts and their contents. The main focus of the analysis was to examine the manuscripts preserving the complete texts of *Hrómundar saga Gripssonar* in Icelandic, therefore the manuscripts containing excerpts and translations were ignored. The distribution of texts appearing frequently alongside *Hrómundar saga Gripssonar* by century has been obtained using XPath queries of the online catalogue *Stories for all times*.

Main Findings

As a result of this research, the hypothesis can be confirmed: generally, texts belonging to one genre appear most frequently in manuscripts with other texts belonging to the same genre. However, an interesting transmission history of *Hrómundar saga Gripssonar* suggests a close association of this saga with the late legendary sagas, and in particular *Bragða-Ölvis saga.* Both *Bragða-Ölvis saga* and *Hrómundar saga Gripssonar* are post-medieval re-workings of older metric versions of the stories (rímur), and for both texts the manuscript AM 601 b 4to (Árni Magnússon Institute, Reykjavík) was suggested as the witness carrying the best text of the saga (Andrews, 1911; Brown, 1946; Hooper, 1934; Hooper, 1932). Even though *Hrómundar saga Gripssonar* appears most frequently with texts classified as late legendary sagas in pre-1800 manuscripts, after 1800 the texts classified as (traditional) fornaldarsögur start to dominate. The late *Bragða-Ölvis saga* dominates the pre-1800 setting, but the distribution changes in the nineteenth century when *Þorsteins saga Víkingssonar*, *Starkaðar saga gamla*, *Friðþjófs saga ins frækna*, and *Hálfs saga Hálfsreka* appear more frequently (as presented on figure below). *Starkaðar saga gamla* is a late-eighteenth century saga written by Snorri Björnson (1710-1803), utilizing traditional legendary motifs of Saxo's *Gesta Danorum* and legendary sagas (Driscoll, 2009:209; Simek and Hermann Pálsson, 1987:331), so its co-occurrence with other legendary sagas starting from the eighteenth century onwards is not surprising. The three remaining texts that started to appear more frequently alongside *Hrómundar saga Gripssonar* in nineteenth-century manuscripts were all published in the same volume of Rafn's *Fornaldarsögur Norðrlanda*, in which *Hrómundar saga Gripssonar* was published (volume II); likewise Friðþjófs saga ins frækna, and *Hálfs saga Hálfsreka* appeared in Björner's edition from 1737 together with *Hrómundar saga Gripssonar.* This shows how printed editions influenced the saga's transmission in the manuscript form. A text, which once showed strong connections to another rímur-based narrative, became detached from its previous setting and gained new, print-influenced context after becoming part of printed editions.

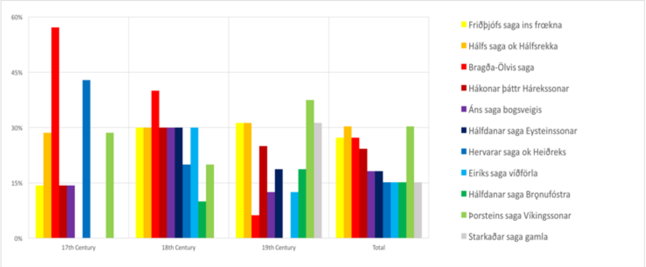


Figure 1. Texts appearing frequently with Hrómundar saga Gripssonar in manuscripts by century

Relevance

The topic of this paper fits in the advertised panel "Quantitative stylistics and philology, including big data and text mining studies," as it employs database quarrying and network analysis of significant amount of data.

Acknowledgements

The author of this paper owes a great debt to Tarrin Wills for merging the data from various databases (fasnl.ku.dk, skaldic.abdn.ac.uk, handrit.is, onpweb.nfi.sc.ku.dk) to obtain the input data for the network analysis, and to Tim Rowbotham, who worked on standardizing the uniform titles and genre classifications in the XML-files in the Stories for all times catalogue.

Bibliography

Guðmundsdóttir, A. (2001). *Úlfhams Saga*. (Stofnun Árna Magnússonar Á Íslandi 53). Reykjavík: Stofnun Árna Magnússonar á Íslandi.

Andrews, A. L. (1911). Studies in the fornaldarsögur Norðurlanda. *Modern Philology*, 8: 527–44.

Björner, E. J. (1737). *Nordiska kämpa dater i en sagoflock samlade om forna kongar och hjältar. Volumen historicum, continens variorum in orbe hyperboreo antiquo regum, heroum et pugilum res praeclare et mirabiliter gestas. Accessit, praeter conspectum genealogicum Svethicorum regum et reginarum accuratissimum etiam praefatio.* Stockholmiae: typis J.L., Horrn. http://books.google.com/books?id=9nZUAAAAYAAJ (accessed 28 January 2016).

Brown, U. (1946). The saga of Hrómund Gripsson and Þorgilssaga. *Saga-Book*, 13: 51–77.

Driscoll, M. J. (2005). Late prose fiction (lygisögur). *A Companion to Old Norse-Icelandic Literature and Culture.* Oxford: Blackwell Publishing Ltd, pp. 190–204.

Driscoll, M. J. (2009). Editing the Fornaldarsögur Norðurlanda. *Á Austrvega, Saga and East Scandinavia, Preprints of the 14th International Saga Conference Uppsala 9th - 15th August 2009.* Gävle: University of Gävle, pp. 207–12.

Gephi. (n.d.) Gephi: The open graph viz platform: https://gephi.org

Hall, A. and Parsons, K. (2013). Making stemmas with small samples, and digital approaches to publishing them: testing the stemma of Konráðs saga keisarasonar. *Digital Medievalist*, 9.

Handrit. (n.d.) Online catalogue handrit.org: http://handrit.org

Harris, J. (1975). Genre in the saga literature: A Squib. *Scandinavian Studies,* 47(4): 427–36.

Hooper, A. G. (1932). “Bragða-Ǫlvis saga” now first edited*. Leeds Studies in English*, 1: 42–54.

Hooper, A. G. (1934). Hrómundar saga Gripssonar and the Griplur. Leeds Studies in English, 3: 51–56.

Kapitan, K. A., Rowbotham, T. and Wills, T. (2017). Visualising genre relationships in Icelandic manuscripts. *Conference Abstracts.* Gothenburg: The University of Gothenburg, pp. 59–62.

Lavender, P. (2015). The Secret Prehistory of the Fornaldarsögur. *The Journal of English and Germanic Philology,* 114(4): 526–51.

Mitchell, S. (1991). *Heroic Sagas and Ballads.* Ithaca and London: Cornell University Press.

Nordisk Forskningsinstitut (n.d). The “Stories for all time” Project. http://fasnl.ku.dk/

Quinn, J. (2006). Interrogating Genre in the Fornaldarsögur: Round-Table Discussion’. *Viking and Medieval Scandinavia*, 2: 276–96.

Rafn, C. C. (1829). *Fornaldarsögur Norðrlanda.* Kaupmannahöfn

Rowe, E. A. (1993). Generic Hybrids: Norwegian ‘Family’ Sagas and Icelandic ‘Mythic-Heroic’ Sagas. *Scandinavian Studies,* 65(4): 539–54.

Rowe, E. A. (2004). ‘Þorsteins þáttr uxafóts, Helga þáttr Þórissonar,’ and the Conversion ‘þættir’. *Scandinavian Studies*, 76(4): 459–74.

Simek, R. and Pálsson, H. (1987). Lexikon Der Altnordischen Literatur*, Die Mittelalterliche Literatur Norwegens Und Islands.* (Kröners Taschenausgabe 490). Stuttgart: Kröner.

Skaldic Project Academy Body. (n.d.) Skaldic project: http://skaldic.abdn.ac.uk/db.php?

University of Copenhagen (n.d.) Ordbog over det norrøne prosasprog Registre: http://onpweb.nfi.sc.ku.dk/mscoll_d_menu.html

Transforming Theater History through Crowdsourced Transcription

Lindsay King
lindsay.king@yale.edu
Yale University, United States of America

Peter Leonard
peter.leonard@yale.edu
Yale University, United States of America

Based on the pilot Ensemble project from NYPL Labs, Ensemble @ Yale is an experiment that aims to transform the archives of Yale’s theater history through crowdsourced transcription. Going beyond OCR of digitized text, Ensemble @ Yale uses human expertise to extract semantic relationships and structured data from digitized programs, enabling future digital scholarship on a rich cultural collection.

The Yale School of Drama is internationally recognized as a leading theater training program, and its associated professional company Yale Repertory Theatre has premiered numerous plays that have gone on to successful productions in New York and elsewhere. The long list of renowned alumni attests to substantial research interest in records of their careers at Yale. Yale University Library's archives of productions on campus, dating back to the School of Drama's origins as a department in 1925, are housed in several disparate special collections units. The vast majority of these records exist solely in print. Even searchable digital finding aids often require that researchers have knowledge of the time period when a particular person was at Yale and might have been in a show. Neither personal names nor show titles are available as entry points in minimally-processed collections like the Yale Repertory Theatre and Yale School of Drama Ephemera Collection, the archive this initial experiment seeks to investigate.

Ensemble asks its users to choose important parts of digitized performance programs, transcribe the text, identify important relationships or characteristics, and verify the work of other users to make the collected data more robust. Crowdsourced transcription relies on individual judgment to pull out features of interest to scholars. A key challenge in crowdsourcing is enabling users to produce the most comprehensive metadata possible, while not making data entry feel like a chore. We strove for a feeling of progress and accomplishment when users complete short tasks, hoping to replicate NYPL Labs' success at attracting transcribers.

Designing the user experience for Ensemble requires a different kind of organization than the original collection/box/folder order, so that the relationship of the crowdsourcing workflows to the special collections materials is both clear and engaging to users. Many transcription systems present the user with random pages one after another. In this case, however, we sought to avail ourselves of a highly-motivated user base with specific domain expertise. For these potential users, navigating directly to programs from specific years they knew well would be a more satisfying experience and also allow the project to leverage their knowledge.

As our work to digitize the programs within one of the archival collections commenced, NYPL Labs shifted focus toward creating a more general crowdsourcing tool. Recognizing the broad applicability of the ideas behind crowdsourced transcription, NYPL Labs partnered with Zooniverse, a "citizen science" organization with a history of crowdsourcing scientific archives, to produce a generalized, reusable toolkit. The resulting software project, called Scribe, was produced with the support of a National Endowment for the Humanities Digital Humanities Implementation Phase grant. With this abstracted transcription engine as a springboard, we began to adapt Scribe to work with multipage theatrical objects in a new version of Ensemble.

Among other adjustments, we are adding a name-authority component to the verification workflow, to enable better searching of names and roles in the database. For production staff roles, a researcher might want to know that a term like "scenery" was used at a specific time, but "set design" and "settings" at other times. They may also want to be able to search for the names of everyone who worked on that enterprise in a certain organization or geographic location. Verification and normalization will be done by librarians "offstage," allowing retention of users' direct transcriptions as data. The resulting structured data will have multiple research applications, from simple database queries for answering reference questions to network analysis and other data visualizations.

In this paper, we will present the project's progress after the public launch of Ensemble @ Yale this spring. Beyond its usefulness for creating structured data from print archives, we are also interested in Ensemble's potential as an outreach tool to engage scholars, alumni, and the general public with archival materials in the library's collections. As an experiment, Ensemble was envisioned as a model that could be expanded to additional repositories at Yale. Beyond that, the proof of concept may inspire similar projects at other institutions that could expand into linked collections of theatrical data across multiple repositories.

Overcoming Data Sparsity for Relation Detection in German Novels

Markus Krug
markus.krug@uni-wuerzburg.de
University of Wuerzburg, Germany

Isabella Reger
isabella.reger@uni-wuerzburg.de
University of Wuerzburg, Germany

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Wuerzburg, Germany

Lukas Weimer
lukas.weimer@uni-wuerzburg.de
University of Wuerzburg, Germany

Nathalie Madarász
nathalie.madarasz@stud-mail.uni-wuerzburg.de
University of Wuerzburg, Germany

Frank Puppe
frank.puppe@uni-wuerzburg.de
University of Wuerzburg, Germany

Introduction

Within the context of social network analysis (SNA) for literary texts the automatic detection of family relations and similar social relations between characters in novels would be an important step for any macroscopic analysis. Manual labeling is rather inefficient since the text snippets that explicitly describe a relation are sparse within the long text documents; therefore we combine two techniques, active learning and distant supervision, which are often used to overcome data sparsity.

Inspired by distant supervision which uses a high quality information resource to support information extraction from other data, we used expert summaries of literary texts, German novels mainly from the 19th century, since relevant text snippets are much more frequent in summaries. Then we applied an uncertainty-based active learning strategy labeling selected sentences from the novels and the complete summaries. The results show that training on summaries and evaluating on data derived from novels yields reasonable results with high precision and low recall similar to humans solving this task.

After a brief discussion of related work in the next section, the data set and the necessary preprocessing for this work are explained in section three. Section four describes our method in detail and shows strengths and weaknesses.

Related work

The challenge of training an algorithm capable of generalizing from a small set of manually labeled data has created a multitude of approaches like active learning and distant supervision. A good survey on active learning is given in (Finn et al., 2003). Usually it starts with a seed set of manually annotated data. A classifier is then trained and new instances that appear to be very different from the current training data are proposed for manual labeling until the quality of the classifier stops improving. Successful algorithms include Multi-instance Multi-label Relation Extraction (Surdeanu et al., 2012).

Another method specifically used for relation detection in newspapers is distant supervision (Mintz et al., 2009): Given some facts (e.g. Michelle Obama is the wife of Barack Obama), usually stored in a database, the aim is to match those facts to the text (e.g. every sentence containing Michelle and Barack Obama indicates that they are married). The training of the classifier is then performed on the pseudo gold data. Even though the idea appears to be simplistic, the results are comparable to those obtained by active learning.

Jing et al. (Jing et al. 2007) successfully applied relation extraction for SNA in an end-to-end manner and reported that most problems were caused by coreference resolution.

Data and preprocessing

We created three datasets from 213 expert summaries, available from Kindler Literary Lexicon Online, and 1700 novels derived from project Gutenberg and annotated relations between characters:

- We split 500 novels into sentences and applied an uncertainty-based active learning strategy (explained below) to iteratively select new examples (in this case full sentences) using a MaxEnt classifier. In total, about 1100 sentences were labeled in this way. This was labeled by annotator 1. (From now on, we refer to this as the novel data set)
- We split our summaries into sentences and applied the same active learning strategy to select new examples, thereby generating about 1300 labeled sentences. They were labeled by annotator 1. (From now on, this is called summaries I)
- Each of the 213 summaries has been manually labeled with all character references, the co-reference chains amongst them and relation annotations for pairs of entities that are explicitly mentioned to be in a relation. They have been labeled by annotator 2. (From now on, we call this summaries II)

The applied active learning strategy started by manually selecting about 20 seed training sentences which were manually labeled with information about relations between character references. The seed examples were chosen by matching a wordlist containing indicative expressions (such as “mother”, “father”, “servant” or “loves”) to the text, to enable the classifier to learn relations from different relation types in an unbiased fashion (which usually changes during training because the underlying distribution of relations is heavily biased towards family relations). On those seed examples we trained a binary Maximum Entropy classifier which was applied to thousands of unlabeled sentences. The sentences were then ranked by uncertainty of the classifier. Uncertainty for a sentence, in our case, was defined by extracting all pairs of character references first, applying the classifier to every pair and then assigning the minimum probability to the sentence. The classifier was retrained on command of the user and the ranking of the unlabeled sentences restarted. By applying this strategy, we observed that the average certainty of a sentence rises with every iteration and decided to stop the manual labeling once there was no sentence with a classifier probability below 60% for the novels and 70% for the summaries (this does not mean we reached saturation in classification gain).

For the labeling, we used a total of 57 hierarchically ordered relation labels, inspired by (Massey et al., 2015) (see figure 1). All these labels relate person-entities with each other, such as “motherOf” or “loves”.

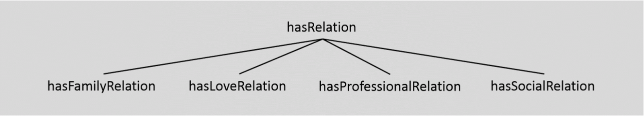
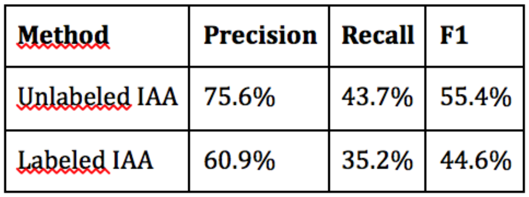


Figure 1: The four main relation types which are further differentiated in 57 relation types in total

The inter-annotator-agreement (IAA) between summary data set I and summary data set II was measured in two ways:

1. A true positive appears when both annotators mark the correct span of the annotation as well as the correct label and the correct arc direction where a correct arc links the two entities in the direction as it is expressed in the text (labeled inter-annotator agreement).
2. A true positive appears when both annotators mark the correct span and arc direction of the relation (unlabeled inter-annotator agreement).

Table 1 gives an overview of the IAA results.



Method	**Precision**	**Recall**	**F1**
Unlabeled IAA	75.6%	43.7%	55.4%
Labeled IAA	60.9%	35.2%	44.6%

Table 1: IAA results, the comparison assumed summaries II as gold and compared summaries I to it.

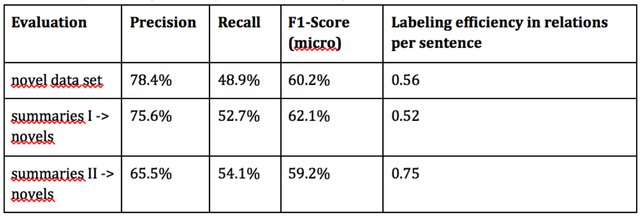
Additionally, we determined 55.5% as the normalized Cohen's Kappa between our annotators. The results for the IAA are surprisingly low (amount of labeled relations in summaries I compared to the relations in summaries II). The reasons are yet unclear and have to be investigated; we assume that one of them is the high variance of possibilities to express social relations. Labeling the complete summaries may also be more difficult because the annotator needs to read the text completely and might use background knowledge to annotate relations which are only implicit in the text.

Method and evaluation

To compare the transfer from summaries to novels, we trained a classifier, specifically a maximum entropy classifier based on boolean features generated from rule templates because previous work has shown that this classifier is superior in classification accuracy compared to kernel machines, pure rule based approaches or other supervised classifiers such as support vector machines (Krug et al., 2017). Training was done on a data set using the annotations as features and the classifier was applied either to test data from the same set or to a different data set resulting in three evaluations:

- A 5-fold cross evaluation within the novel data set.
- Training on the snippets of the summaries (summaries I or summaries II) and evaluation on the novel data set.

Table 2 shows the result of this experiment for the in-data and cross data evaluation of the relation detection component.



Evaluation	**Precision**	**Recall**	**F1-Score (micro)**	**Labeling efficiency in relations per sentence**
novel data set	78.4%	48.9%	60.2%	0.56
summaries I -> novels	75.6%	52.7%	62.1%	0.52
summaries II -> novels	65.5%	54.1%	59.2%	0.75

Table 2: The results of a 5-fold in-data set evaluation for both of the data sets and the results for a cross-data set evaluation. Each number represents a micro-average score, i.e. we count every true-positive, false-positive and false-negative in a document and calculate the average scores based on these quantities. We choose the micro score since the label set is rather unbalanced between classes. The efficiency of an approach is measured by calculating number of relations / number of sentences.

Results that are very similar to working directly on the novel (60.2% F1) are achieved by using the model trained on the extracted sentences from the summaries to retrieve information about character relations in the novels (62.1% resp. 59.2%). Since our test data is generated by active learning and only the most difficult examples were chosen for labeling, we expect our results to be a lower bound compared to data in complete novels.

If we use a model trained on the complete summaries, we experience a drop in precision. This drop was to be expected, since the amount of additional labeled relations in the novels is high according to the IAA results (this manifests in the low recall in table 1) as well as can be seen in the labeling efficiency. Altogether, the quality and efficiency of using a classifier trained on summaries are comparable to training on the novels directly based on our data.

Summary

We presented an approach to increase labeling efficiency for relation detection in German novels by transferring knowledge from summaries to novels. It could be shown that using the summaries as trainings data will achieve similar results to using the novels, but the summaries are much shorter and relevant sentences are much more frequent. The inter-annotator agreement for this task is also relatively low which may point to an explanation for the comparatively low results of the automatic approach.

Bibliography

Finn, A., and Kushmerick, N. (2003). "Active learning selection strategies for information extraction." *Proceedings of the International Workshop on Adaptive Text Extraction and Mining (ATEM-03).*

Jing, H., Kambhatla, N. and Roukos, S. (2007). "Extracting social networks and biographical facts from conversational speech transcripts." *Proceedings of the 45th Annual Meeting of the*

Association for Computational Linguistics, Prague: ACL, pp. 1040-48.

Krug, M., Wick, C., Jannidis, F., Reger, I. Weimer, L., Madarász, N. and Puppe, F. (2017). “Comparison of Methods for Automatic Relation Extraction in German Novels.” 4. *Tagung Digital Humanities im deutschsprachigen Raum.* Bern: DHd, pp. 223-26.

Massey, P., Xia, P., Bamman, D. and Smith, N. A. (2015). “Annotating character relationships in literary texts.” arXiv preprint: arXiv:1512.00728.

Mintz, M., Bills, S., Snow, R. and Jurafsky, D. (2009). “Distant supervision for relation extraction without labeled data.” *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Singapore: ACL, pp. 1003-11.

Surdeanu, M., Tibshirani, J., Nallapati, R. and Manning, C. D. (2012). “Multi-instance multi-label learning for relation extraction.” *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* Jeju Island, Korea: ACL, pp. 455-65.

La production de l’espace dans l’imprimé d’Ancien Régime: le cas de la Gazette

François Dominic Laramée
fdl@francoisdominiclaramee.com
Université de Montréal, Canada

Quelle image mentale du monde un Français de l’époque moderne pouvait-il se tracer au contact des journaux et des livres? Comment caractériser le « message géographique » transmis par les ouvrages savants, les périodiques, les descriptions et les récits de voyage du XVIII^{e} siècle — et peut-être discerner leur influence sur les événements historiques ?

L’étude d’une telle problématique, qui constitue le cadre de ma thèse de doctorat, implique de faire appel à un corpus vaste et diversifié qui doit être examiné à la fois à l’aide de méthodes numériques de traitement de la langue naturelle et par une lecture intensive de documents ciblés. Mais comment appliquer la lexicométrie, la fouille de textes et l’apprentissage machine à des textes en français du XVIII^{e} siècle, pour lesquels il n’existe pas de modèle de langage approprié dans les logiciels d’analyse courants ? Et comment adapter des algorithmes conçus pour des sources numériques récentes et de très bonne qualité à des documents endommagés par les siècles, parfois océrisés à partir de microfilms alignés de façon imprécise, ou même handicapés dès leur création par une typographie irrégulière qui mystifie les outils d’océrisation ?

Cadre théorique

Depuis les années 1980, le tournant géographique inspiré par les travaux du philosophe Henri Lefebvre (Lefebvre, 2000) a démontré que l’espace est une construction sociale. Pour les Français du XVIIIe siècle, cette construction passait le plus souvent par la lecture, seule source de connaissance couramment disponible au sujet d’espaces lointains. Comment peut-on utiliser les concepts d’espace et de lieu (Tuan, 2006), de co-présence et de mobilité (Lévy, 1999; Lussault, 2007) pour reconstruire l’imaginaire spatial suscité au sein de communautés de lecteurs par l’accès simultané aux mêmes textes (Anderson, 2006), en particulier ceux des journaux et périodiques?

Corpus et méthodologie

Cette présentation examinera les résultats obtenus lors d’une analyse d’un corpus tiré de la Gazette (renommée Gazette de France en 1762), principal périodique de nouvelles sous l’Ancien Régime (Feyel, 2000). Une version numérisée de la Gazette est disponible sur Gallica, l’archive en ligne de la Bibliothèque nationale de France ; l’état de conservation des documents imprimés à partir desquels cette version a été constituée est cependant inégal, et par conséquent le taux de succès estimé à l’océrisation varie entre 99 % et 76 % ou même moins. Les fichiers .txt issus de l’océrisation ne peuvent donc pas être employés tels quels : non seulement les formes (chaînes de caractères représentant des mots) sont-elles fréquemment endommagées, mais des informations aussi importantes que les limites séparant deux articles et les lieux d’origine de ceux-ci sont également sujettes à un taux d’erreurs inacceptable. La retranscription manuelle du corpus, formé de dizaines de milliers de page en PDF, aurait quant à elle entraîné un coût d’acquisition déraisonnable. Quant aux méthodes usuelles de correction automatique des erreurs d’océrisation (Lopresti, 2009), elles se sont révélées peu efficaces dans ce contexte, ne permettant de réduire le taux d’erreur effectif que de moins de 0,1% — et le plus souvent dans des segments du corpus sans lien direct avec les questions de recherche étudiées.

Pour traiter ce corpus, il a donc fallu faire appel à une méthode itérative, où le choix de questions de recherche auxquelles répondre a déterminé les éléments du corpus qu’il fallait reconstruire et où les résultats de l’examen d’une version transitoire du corpus a guidé le choix des questions de recherche pour l’étape suivante. Cette méthode repose sur l’identification dans le corpus, à l’aide de l’algorithme de Levenshtein (Crump, 2014), de formes potentiellement produites par une reconnaissance incorrecte d’un certain nombre de mots-clés choisis en fonction d’une question de recherche; sur l’inspection visuelle de ces formes candidates pour éliminer de la liste celles qui correspondent manifestement à d’autres mots de la langue française que les mots-clés recherchés; et sur l’extraction semi-automatisée de métadonnées pertinentes compte tenu de la question de recherche étudiée, à partir du texte océrisé et d’une inspection visuelle du document PDF d’origine. En pratique, le taux de faux positifs obtenus en détectant toutes les formes dont la distance de Levenshtein par rapport aux mots-clés est de 3 ou moins dépasse les 95%, mais la sélection visuelle des candidates véritablement prometteuses permet d'augmenter le nombre d'occurrences utilisables pour une analyse ultérieure de 25% à 30%, et ainsi d'assurer une meilleure couverture des éléments pertinents du corpus que ce qui aurait été possible autrement. (Notons que ces occurrences récupérées incluent non seulement les résultats d'erreurs d'océrisation mais aussi des orthographes inusitées des mots-clés recherchés.)

La présentation tracera les grandes lignes de ce processus, des résultats obtenus avec la Gazette, des éléments de la méthode qui se sont montrés généralisables à d'autres corpus bruités, et des limites que la prudence impose à la fois aux questions de recherche auxquelles il convient d'appliquer une telle méthode et aux conclusions que l'on peut en tirer. Afin d'augmenter le niveau de confiance envers les résultats, une multiplicité de méthodes numériques ont été appliquées aux textes et aux métadonnées, l'utilisation d'un seul algorithme, toujours problématique (Schmidt, 2013) étant particulièrement suspecte dans un contexte où la fiabilité des données laisse à désirer. Ces multiples méthodes incluent le partitionnement (Chen et al., 2004), la cartographie numérique, divers décomptes et l'étude des cooccurrences lexicales — soigneusement contrôlée par une inspection visuelle afin d'éliminer les effets de bord causés par l'absence d'un modèle de langage approprié pour le français du XVIIIe siècle — avec le logiciel de textométrie TXM (Heiden et al., 2010). Seuls les résultats à la fois cohérents entre les différentes méthodes et trop flagrants pour être expliqués par un accident de répartition du bruit dans les textes d'origine ont été conservés pour communication.

Résultats

Les premiers résultats portent sur la représentation de l'Amérique et du monde colonial au cours de la période entre 1740 et la fin de la Guerre de Sept ans. Il a notamment été possible de démontrer que l'immense majorité des articles de presse mentionnant les colonies provenaient de Londres ou de la péninsule ibérique plutôt que de la France elle-même et qu'ils présentaient le phénomène colonial d'un point de vue étranger ; que les colonies britanniques et le Brésil occupaient une place beaucoup plus importante que les colonies françaises dans l'imaginaire spatial construit par la Gazette ; et que la sous-représentation du monde colonial français dans les textes, et en particulier celle de la Nouvelle France continentale, était exacerbée en temps de paix, où le Canada devenait pratiquement invisible. À la lecture de ces résultats, il est permis de se demander si, du point de vue d'un lecteur de la Gazette, le moment colonial français en Amérique n'aurait pas semblé chose du passé, bien avant la signature du traité de Paris qui a consacré la cession du Canada à la Couronne britannique. Des recherches sur la co-présence de différents toponymes au sein des mêmes articles, des thèmes associés à ces toponymes et à la distance imaginaire représentée par les fréquences de mentions de lieux dans la Gazette sont en cours et leurs résultats pourront être intégrés à la présentation.

Bibliographie

Anderson, B. (2006). Imagined Communities: Reflections on the Origin and Spread of Nationalism, revised edition. London: Verso.

Chen, J., Ching, R. and Lin, Y. (2004). "An Extended Study of the K-Means Algorithm for Data Clustering and Its Applications." The Journal of the Operational Research Society, 55(9): 976-987.

Crump, J. (2014). "Generating an Ordered Data Set from an OCR Text File." Programming Historian, http://programminghistorian.org/lessons/generating-an-ordered-data-set-from-an-OCR-text-file

Feyel, G. (2000). L'annonce et la nouvelle: la presse d'information en France sous l'Ancien Régime, 1630-1788. Oxford: Voltaire Foundation.

Heiden, S., Magué, J-P. et Pincemin, B. (2010). "TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement." Proceedings of 10th International Conference on the Statistical Analysis of Textual Data — JADT 2010. Rome: Edizioni Universitarie di Lettere Economia Diritto, vol. 2, pp. 1021-1032.

Lefebvre, H. (2000). La production de l'espace, 4e édition. Paris: Anthropos.

Lévy, J. (1999). Le tournant géographique: penser l'espace pour lire le monde. Paris: Belin.

Lopresti, D. (2009). "Optical Character Recognition Errors and Their Effects on Natural Language Processing." International Journal on Document Analysis and Recognition, 12 (3): 141-51.

Lussault, M. (2007). L'homme spatial: la construction sociale de l'espace humain. Paris: Seuil.

Schmidt, B. (2013). "Words Alone: Dismantling Topic Models in the Humanities." Journal of Digital Humanities, 2(1): 49-65.

Tuan, Y. (2006). Espace et lieu: la perspective de l'expérience, Gollion: Infolio.

Toward Reproducibility in DH Experiments: A Case Study in Search of Edgar Allan Poe's First Published Work

Mark D. LeBlanc
mleblanc@wheatoncollege.edu
Wheaton College, United States of America

Summary

Reproducing experimental results is a hallmark of empirical investigation and serves both to verify and inspire. This paper is a call for more systematic documentation of computational stylistic experiments. Publishing only summaries of the methods and results of empirical work is an artifact of traditional print media. To facilitate experimental

reproducibility and to help the growing community who wish to learn how to apply computational methods and subsequently teach the next generation of scholars, the publication of results must include (i) access to the digitized texts, (ii) a clear workflow and most essentially (iii) the source code that led to each and all of the experimental results. By way of example, we present the steps and process in a GitHub repository for computationally probing the unknown and contested authorship of an 1831 short story entitled "A Dream" as we seek evidence if this work is similar to other attributed works by Edgar Allan Poe. The entire framework is intended as a pedagogical jumpstart for others, especially those new to computational stylometry. If Poe did write the story, it would be his first published work.

Introduction

As the Digital Humanities gains access to a wide array of digitized corpora and matures to a discipline that creatively defines new methods for computationally close and distant readings, a growing gap has emerged between those who apply sophisticated programming, e.g., Stylo In R (Eder *et al.*, 2016) and those who are new to the game and need an introduction to the field. Typical of the community spirit in DH, significant efforts are underway to bridge this gap, including web-based tools for entry-level exploration including Voyant Tools (Sinclair and Rockwell, 2016) and Lexos (Kleinman *et al.*, 2016) and domain-specific introductions to programming, including Jockers' text (2014) and the Programming Historian (Crymble *et al.*, 2016). This paper attempts to narrow the gap by encouraging both sides to document their experimental methods more fully to embrace previous calls for the replication of experimental methods (Rudman, 2012 *et al.*) and thereby teach effective practices by "leaving a trail" of experimental methods that enable others to execute and extend.

A Good Mystery: Towards Reproducibility

A GitHub repository or "repo" offers a workflow that explores whether an 1831 story published under the attribution of only 'P' might have been written by Edgar Allan Poe. If so, it would be Poe's first published work. In addition to sharing a set of analytical methods applied in this experiment, the broader methodological-pedagogical goals are two-fold: (i) the dissemination of data and code should be championed as a cornerstone of DH research, thereby facilitating the replication of results and (ii) to share a workflow so that others may apply similar analyses to their texts of interest.

The workflow is stored as a set of numbered folders containing the texts *and* scripts (code) needed to complete each step. The workflow includes: collecting texts, the preprocessing, tokenization, and culling decisions made, unsupervised cluster analyses (k-means, hierarchical-agglomerative, bootstrap consensus tree), and supervised classification methods using Stylo in R's Delta, SVM, and NSC models. Each step represents scaffolding for a "teachable moment" with materials provided so faculty can more easily use them with students.

Scrubbing, Tokenization, Cutting, and Culling

Lexos, a web-based, open-source workflow of tools (Kleinman, *et al.*, 2016) was used to upload texts and "scrub" them by applying the following options: (i) convert words to lowercase, (ii) all punctuation was removed, (iii) however, a single word-internal hyphen and word-internal apostrophes were kept, and (iv) all digits were removed. Each individual word is considered as its own token. Larger stories were segmented ("cut") into pieces. We experimented with various culling options, e.g., keeping only the most frequent words that appear in each text at least once.

Cluster Analysis

As a set of initial probes, we compared the contested story "A Dream" to (i) other stories attributed to Poe and (ii) mixed in with stories by other contemporaries. In the repo, we share four variations using cluster analysis:

1. K-means clustering on only Poe's stories (using Lexos)
2. Hierarchical agglomerative clustering on only Poe's stories (uses a Python sklearn module and a script to convert the cluster to ETE and Newick formats)
3. K-means clustering when all stories by each author are concatenated together (Lexos)
4. Bootstrap Consensus Tree (using Stylo in R).

The result from the Bootstrap Consensus Tree is shown in Figure 1. Of interest is that each author's stories cluster consistently together (with the exception that Bird's initial section of "Sheppard Lee" and his "Calavar" are found in different clades, at six and eight o'clock). "A Dream" clusters with the smaller Poe texts. As you'll see, we couldn't resist tossing in the four stories sometimes attributed to Edgar's brother Henry ("Monte Video", "A Fragment", "The Pirate", and "Recollections"). These four stories are found within the cluster of Poe's known works (*c.f.* Collins, 2013).

A series of cluster analyses often serves well as a preliminary exploration, especially for scholars who are new to this game. Some of the file sizes are very small (*e.g.*, one-half of the Poe stories in this corpus have fewer than 2000 words) and when strict culling is enforced (top-N words that appear at least once in each segment), the available set of words is reduced to only 38 when dealing with "A Dream" and the other eighteen Poe stories. That noted, these exploratory investigations shed some light on why some scholars consider that Poe's "first published tale may have been 'A Dream'" (Silverman, 1991, p87).

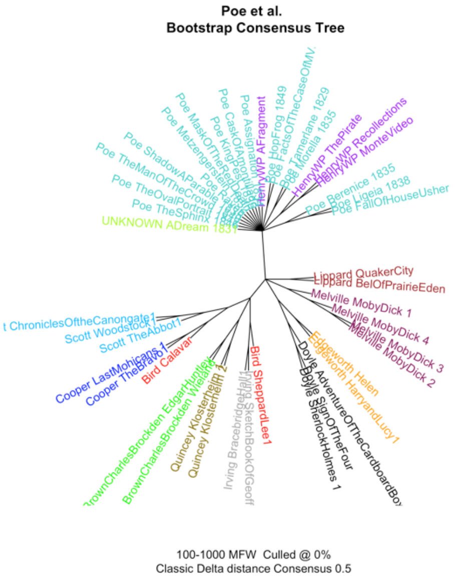


Figure 1. Using Stylo in R Bootstrap Consensus Tree (BCT) showing "A Dream" consistently clustering with other Poe stories. The BCT aggregates results over multiple cluster analyses and shows those texts that satisfy a consensus number of the individual trials. Using 12 different authors and at least two texts by each author for a total of 46 stories, Stylo formed clusters of the texts for the following frequency bands when using the most-frequent words: 100 to 1000 MFW.

Classification

Three classification models differentiated authorial writing style as implemented in Stylo in R. We scripted in R alongside Stylo to test "A Dream" over N-trials (N=10, 100) using a random selection of files for training sets in each trial. At least one text from each author is also included in the test set for each trial. A follow-up Python script parses the collected results to build confusion matrices for each author to provide metrics on how well the models predict each author's works. The most-frequently occurring, top-40 words (MFW, 1-grams) that appear in all the texts at least once were used.

		Confusion Matrix values for all Poe Stories			
Model	Attributions of "A Dream" to Poe	True+	True-	False+	False-
Delta	9	13	200	0	7
NSC	10	16	170	30	4
SVM	7	14	198	2	6

Table 1: Attributions of the contested story "A Dream" over ten (10) trials with "A Dream" and another randomly selected Poe story in the test set in every trial. Confusion matrix values for results of testing Poe texts over all trials provide overall measures of model effectiveness. In the three cases where "A Dream" was attributed to a different author, Poe was ranked second.

Summary

We offer a start to an exploration to collect evidence as to whether Poe may have written the 1831 story "A Dream" (*c.f.,* Schöberlein (2016) who used the most frequent character 3-grams and attributed the story to Poe using Delta, but not so when using NSC nor SVM models). Evidence and methods aside, a GitHub repo provides a framework to share experimental workflows in a spirit similar to Jupyter notebooks, as well as one that facilitates both reproducible results and opportunities for subsequent contributions.

Notes

Forming an appropriate corpus is hard: thanks to Sam Coale, Ryan Cordell, Cary Gouldin, David Hoover, Shirrel Rhoades, and Ted Underwood. Four undergraduates: Weiqi Feng, Alec Horwitz, Jingxian Liu, and Khaled Sharafaddin worked with us on this problem. Thanks to Maciej Eder for his help with Stylo in R.

Bibliography

Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Taparata, E., Visconti, A., and Wieringa, J., eds. (2016). *The Programming Historian*. 2nd ed.. Web: http://programminghistorian.org/.

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, 16(1): 107-121.

GitHub repository: A Good Mystery. . https://github.com/WheatonCS/aGoodMystery

Jockers, M. (2014). *Text Analysis with R for Students of Literature.* Springer, New York.

Kleinman, S., LeBlanc, M.D., Drout, M., and Zhang, C. (2016). Lexos v3.0. Web: http://lexos.wheatoncollege.edu.

Rudman, J. (2012). The State of Non-Traditional Authorship Attribution Studies -- 2012: Some Problems and Solutions. *English Studies*, v93(3), 259-274.

Schöberlein, S. (2016). Poe or Not Poe? A Stylometric Analysis of Edgar Allan Poe's Disputed Writings. *Digital Scholarship in the Humanities*, July 24, 2016.

Silverman, K. (1991). *Edgar A. Poe: Mournful and Never-Ending Remembrance*. HarperCollins, New York.

Ecriture et visualisations numériques

Christophe Leblay
christophe.leblay@utu.fi
Université de Turku, Finland

Gilles Caporossi
gilles.caporossi@hec.ca
HEC Montréal, Canada

Depuis les années 80, plusieurs méthodes pour analyser le processus d'écriture ont été utilisées (Miller & Sullivan, 2006). L'outil principal pour analyser le processus d'écriture est le fichier d'enregistrement, appelé *log*, qui contient de façon exhaustive et détaillée l'ensemble des opérations effectuées par le scripteur lors de la rédaction d'un texte (Sullivan & Lindgren, 2014). Les données qui y sont emmagasinées sont considérables, et lorsqu'elles ne sont pas préalablement traitées, elles sont hostiles à l'analyse humaine (Caporossi & Leblay, 2014). Ce traitement préalable nous semble donc déterminant dans l'accès aux données scripturales enregistrées. Autrement dit, il est impossible d'exploiter les programmes d'enregistrement de l'écriture (dits aussi de *temps réel*), sans procéder, au préalable, à un *redéploiement visuel* des données obtenues. Les structures sous-jacentes des données ainsi représentées sont généralement plus propices à l'analyse que les données brutes inexploitables. Plusieurs étapes sont alors nécessaires avant même de pouvoir utiliser les techniques d'analyse à proprement parler. Le processus simplifié par lequel les données sont acquises, traitées et analysées se résume ainsi : a) *enregistrement*, b) *partage*, c) *tri*, d) *recherche*, e) *représentation visuelle des données* et f) *analyse* (Manyika*, et al*., 2011). Plusieurs de ces actions ont déjà été étudiées de manière individuelle et ont été l'objet de création de logiciels visant à enregistrer le processus d'écriture et à le représenter dans l'une ou l'autre de ses dimensions (Caporossi & Leblay, 2011). Ces logiciels sont tous différents et généralement destinés à un projet de recherche particulier (Sullivan & Lindgren, 2014). Certains sont conçus pour traiter l'information et la retransmettent ensuite à l'utilisateur de façon simplifiée, sous la forme d'une représentation visuelle. L'utilisation de représentations graphiques de données, appelées de façon plus générale *visualisations*, consiste à explorer et essayer de comprendre les grands ensembles de données (Yau, 2011) Elles permettent notamment d'identifier des tendances, structures, irrégularités et relations entre les données sur une certaine période temporelle (Minelli, et al., 2013, p. 110). Le format compact de la visualisation agrège les données et utilise les capacités cognitives de l'humain (Blanchard, 2005) Le but principal de l'utilisation de ces images représentant le processus scriptural est de laisser l'œil trouver des structures sous-jacentes parmi les données (Tory & Moller ,2004).

Il existe deux grands modèles de représentations qui prétraitent les données de manière à pouvoir appliquer des techniques d'analyse linguistique de données. Le tout premier, attaché aux travaux réalisés par une linguistique cognitive, est celui qui a été développé dans le cadre des *Systèmes d'Information Géographique*, ou SIG ; ont utilisé ce mode de représentation principalement les études suivantes : *LS Graph* (Lindgren & Sullivan, 2002 ; Leijten, *et a*l., 2006), *Genèse du texte* (Doquet-Lacoste, 2003), *GIS Graph* (Lindgren, *et al*., 2007), *Timeline* (Wengelin, 2009) et *Inputlog* (Leijten, *et al*., 2013). Le second, celui que nous proposons, en réaction au premier, a été développé en associant, de manière interdisciplinaire, la linguistique génétique et la théorie mathématique des graphes.

La représentation par les graphes (des nœuds reliés entre eux par des liens, ou arcs) permet de mettre davantage en relief l'aspect dynamique de l'écriture (cf. figure 1). Celle-ci est orientée sur la chronologie du processus d'écriture (Leblay, 2011). Un nœud (nommée *cellule* dans notre travail) représente la production d'une suite ininterrompue de frappes au clavier (caractères et espaces).

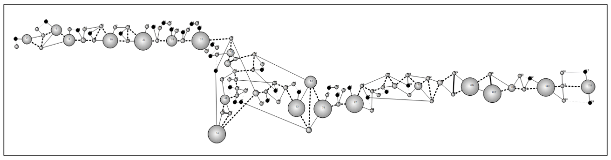


Figure 1. Visualisation globale par graphe : l'écriture experte (nuance de gris : en sombre = une suppression ; en clair = un ajout)

Ainsi, si l'écriture consistait en une suite ininterrompue de frappes de caractères d'un début jusqu'à une fin, un texte serait visuellement représenté par une seule et unique *cellule* dont la taille dépendrait uniquement du nombre de caractères et d'espaces produits. Or, il existe des retours et des pauses dans ce qui est déjà écrit. Les retours dans le texte sont ainsi marqués: la cellule se divise dès l'instant que la continuité topographique est rompue, bien qu'un lien perdure pour matérialiser le lien temporel (couleurs/nuance de gris et épaisseur du trait). Deux cellules se lient donc quand un lien topographique est créé, le tout au gré des écritures et réécritures.

L'une des particularités de ce travail de visualisation est de bien gérer la transformation et les mouvements du texte grâce aux nœuds représentant des parties de texte (ajoutés, supprimés) reliés ensemble par des liens (ou arcs) définissant leur relation, soit chronologique ou spatiale(Southavilay, Yacef, Reimann, & Calvo, 2013) est possible de voir le contenu de ces nœuds. Telle que mentionnée par Caporossi et Leblay (2011), cette représentation en couleurs, ou en nuances de gris, montre l'aspect temporel de la rédaction, soit le moment exact où le scripteur a effectué chacune des opérations représentées, en lien avec celles qui la précédent comme avec celles qui la suivent. Les possibilités d'analyse de ces structures ont déjà été étudiées dans le contexte du processus de l'écriture (Leblay & Caporossi, 2014 ; Caporossi & Leblay, 2015).

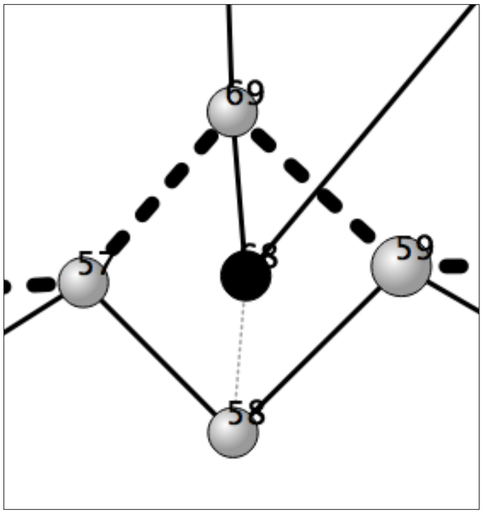


Figure 2. Sous-graphe. Le remplacement

La mise en évidence de sous-graphes particuliers (cf. figure 2) représentant les *patterns* des opérations les plus fréquentes (ajout vs. insertion, suppressions immédiate et différée, remplacement), a déjà été réalisée. L'identification de sous-graphes est utile dans l'analyse globale du graphe représentant le temps de l'écriture (Caporossi & Leblay, 2014)

C'est dans ce cadre que nous proposons 1) un programme dédié, *GenographiX*, pour mettre en évidence tout le travail génétique de réécriture, et 2) un corpus numérique de 10h 20 minutes d'enregistrement ; ce corpus est composé de deux tâches successives, la première de nature narrative (« depuis cette aventure… »), la seconde de nature argumentative (« Qu'est-ce qui est important pour vivre ensemble ? »). Il a été fait le choix de consignes à même de fonctionner auprès de publics très variés (âge, langue maternelle, niveau d'expertise). Précisons que la première reprend un protocole qui a déjà fait l'objet de recueil de corpus de textes auprès d'enfants et de futurs enseignants en formation (Garcia-Debanc et Bonnemaison, 2014), protocole proposé au départ par Charolles (1988) dans le cadre d'études de la cohérence textuelle. L'analyse des processus d'écriture enregistrée trouvera ainsi un point d'appui dans ces études consacrées aux produits.

Trois grands axes de recherche émergent alors : (1) le rôle des *opérations génétiques d'écriture-réécriture* qui sous-tendent toute activité scripturale sur papier comme sur écran (Fenoglio, 2012 ; Grésillon, 1994), (2) les différentes chronologies exhaustives qui caractérisent les écritures experte (vs. novice) et universitaire (vs. scolaire), (Bécotte *et al.*,2016 ; Leblay *et al.*, 2015), et (3) l'impact de la *visualisation* de cette chronologie sur la description et la reproduction de ces écritures dans le cadre de l'enseignement de l'écriture (Doquet & Leblay, 2014 ; Plane *et al.*, 2010). Les résultats que nous proposons, dans cette présentation, concernent particulièrement le phénomène de cohésion textuelle, tel qu'il apparaît, non pas seulement dans un texte produit, mais bien dans son déroulement chronologique. Apparaissent alors des retours ponctuels (qualifiés habituellement d'*erreurs, d'écarts, de dysfonctionnements,* tout simplement de *variantes*) pris sur le vif, en fonction de différents degrés d'expertise scripturale. Il s'agit de traces génétiques, rendues visibles, de phénomènes de retours dans la construction du sens (cohérence/cohésion) qui n'ont été jusqu'à présent observés et décrits uniquement dans des études basées sur des textes produits.

Ces trois axes nous permettent de souligner la place des visualisations numériques proposées par la génétique textuelle contemporaine dans des dispositifs conçus pour la formation à l'écriture : ces visualisations participent pleinement au travail de réflexivité, à celui de l'articulation entre savoirs théoriques et expérimentations de terrain, et enfin, à celui de la formation de praticiens-chercheurs (Brunel & Rinck, 2017).

Bibliographie

Bécotte, H.S., Caporossi, G., Hertz, A. & Leblay, C. (2017). "Writing and rewriting: Keystroke logging's colored numerical visualization." In Sullivan K. P. H. & Lindgren E. (eds), *Observing writing: logging handwriting and computer keystrokes*. Leyde: Brill Academic Publishers. À paraître.

Blanchard, F. (2005). *Visualisation et classification de données multidimensionnelles application aux images multicomposantes*. Presses de l'Université de Reims Champagne Ardenne.

Brunel, M. & Rinck, F. (2016). "Comment former des enseignants spécialistes de l'écriture et de son enseignement ?" *Pratiques 171-172*, L'écriture professionnelle. https://pratiques.revues.org/3197

Caporossi, G. & Leblay, C. (2015). "A graph theory approach to online writing data visualization." In Cislaru, G. (ed.) *Writing(s) at the Crossroads: The Process-Product Interface*. Amsterdam: John Benjamins, pp. 171-181.

Caporossi, G., & Leblay, C. (2014). "Outils de visualisation de données enregistrées." In Leblay, C. & G. Caporossi, G. (éds.), *Temps de l'écriture: enregistrements et représentations*. Louvain-la-Neuve: Academia, pp. 147-166

Caporossi, G., & Leblay, C. (2011). "Online Writing Data Representation : A Graph Theory Approach." *Lecture Notes in Computer Sciences 7014*, pp. 80-89

Charolles, M. (1988). "La gestion des risques de confusion entre personnages dans une tâche rédactionnelle." *Pratiques* 60, pp. 75-97.

Doquet-Lacoste, C. (2003). *Étude Génétique de l'Écriture sur Traitement de Texte d'Élèves de Cours Moyen 2, Année 1995-1996.* Paris: Université Sorbonne nouvelle

Doquet, C. & Leblay, C. (2014). "Temporalité de l'écriture et génétique textuelle: vers un autre métalangage?" In Neveu F. *et al.* (éds.), *Actes numériques du 4ème Congrès Mondial de Linguistique Française*, CMLF, Berlin, 19-23.07.2014. En ligne : http://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01204.pdf

Fenoglio, I. 2012 (Ed). *Genesis* 35, Le geste linguistique.

Garcia-Debanc C. et Bonnemaison K. (2014). "La gestion de la cohésion textuelle par des élèves de 11-12 ans : réussites et difficultés." In Neveu F. *et al.* (éds.), Actes du Congrès Mondial de Linguistique Française CMLF, 961-976. En ligne : http://www.shs-conferences.org/articles/shsconf/pdf/2014/05/shsconf_cmlf14_01349.pdf

Grésillon A. (1994). *Eléments de critique génétique : lire les manuscrits modernes*. Paris, Presses Universitaires de France.

Leblay, C. (2011). *Le Temps de l'Écriture. Genèse, durée, représentations*. En ligne : https://www.jyu.fi/ajankohtaista/arkisto/2011/11/tiedote-2011-11-04-10-14-59-722468

Leblay, C., Caporossi, G., Foucambert, D. & Libersan, L. (2015). "Ecriture & réécriture en situation de travail. Visualisation de pratiques expertes." In Beaudet, C. & Rey, R. (éds.), *Actes du Colloque International L'écriture experte : enjeux sociaux et scientifique*s, Sherbrooke, 13-14.06.2013. Aix-en-Provence: Presses universitaires de Provence, pp. 115-130.

Leijten, M., & Van Waes, L. (2006). "Inputlog : New Perspectives on the Logging of On-Line Writing Processes." In Lindgren K. P. (ed.), *Computer Keystroke Logging and Writing*. Elsevier. Pp. 73-94

Leijten, M., & Van Waes, L. (2013). "Keystroke Logging in Writing Research : Using Inputlog to Analyze and Visualize Writing Processes." *Written communication 30*(3), pp. 358-392.

Lindgren, E., & Sullivan, K. P. (2002). "The LS Graph : A Methodology for Visualizing Writing Revision." *Language Learning, 52*(3), pp. 565-595

Lindgren, E., Sullivan, K. P., Lindgren, U., & Spelman Miller, K. (2007). "GIS for Writing: Applying Geographical Information Systems Techniques to Data Mine Writings' Cognitive Processes." In Rijlaarsdam, G., Galbraith, D., Torrance, M. & Van Waes, L. (eds.), *Writing and Cognition*. Amsterdam: Elsevier. Pp. 83-96

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: the next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Miller, K. S., & Sullivan, K. P. (2006). "Keystroke Logging: An introduction." In Sullivan, K. P. & Lindgren, E. (eds.), *Computer keystroke logging and writing*. Oxford: Elsevier. Pp. 1-10.

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big data, big analytics : Emerging business intelligence and analytic trends for today's businesses.* Indianapolis: Wiley Publishing.

Plane, S., Alamargot, D. & Lebrave, J.-L. (2010). "Temporalité de l'écriture et rôle du texte produit dans l'activité rédactionnelle." *Langages* 177, pp. 7-28.

Southavilay, V., Yacef, K., Reimann, P., & Calvo, R. A. (2013). "Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models." *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 38-47.

Sullivan, K. P., & Lindgren, E. (2014). "La révision en production écrite enregistrée." In Leblay, C. & Caporossi, G. (éds.), *Temps de l'écriture: enregistrements et représentations*. Louvain-la-Neuve: Academia. Pp. 71-92.

Tory, M., & Moller, T. (2004). "Human factors in visualization research." *IEEE Transactions on visualization and computer graphics, 10*(1), pp. 72-84.

Wengelin, A., Torrance, M., Holmwvist, K., Simpson, S., Galbraith, D., Johansson, V., & Johansson, R. (2009). "Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production." *Behavior Research Methods, 41*(2), pp. 337-351.

Yau, N. (2011). *Visualize this: the flowing data guide to design, visualization and statistics.* Indianapolis: Wiley Publishing.

Smelly London: visualising historical smells through text-mining, geo-referencing and mapping

Deborah Leem
d.leem@wellcome.ac.uk
Wellcome Trust, United Kingdom

Overview

Wellcome Collection is one of the world's major resources for the study of health and histories. Over the past few years Wellcome have been developing a world-class digital library by digitising a substantial proportion of their holdings. As part of this effort, approximately 5,500 Medical Officer of Health (MOH) reports for London spanning from 1848-1972 were digitised in 2012. Currently Wellcome holds the most comprehensive digital collection of the London MOH reports. Since September 2016 Wellcome have been digitising 70,000 more reports covering the rest of the United Kingdom (UK).

The MOH reports were published annually by the Medical Officers of Health employed by local authorities across the UK. These reports provided vital statistics and a general statement on the health of the population. MOH reports concentrated on reporting infectious diseases and resolving the problems as well as covering other areas of social responsibilities. (Chave, 1987) They have been long regarded as an important source for the 19th and 20th century history of Public Health and stem from reaction to infectious disease in the mid-19th century. Although there were attempts at standardisation, the reports display each MOH's interest, idiosyncrasies and particular strengths. Therefore, they also provide a particular perspective on the everyday lives of Londoners over several generations. No digital techniques have yet been applied successfully to add value to this very rich resource.

As part of the Smelly London project, the OCR-ed text of the MOH London reports has been text-mined using the Python programming language. Through text mining we produced a geo-referenced dataset containing smell categories for visualisation to explore the data. At the end of the Smelly London project the MOH smell data will also be available through other platforms such as Good City Life and Layers of London. This will allow the public and other researchers to compare smells in London from the 19th century to present day. This has the further potential benefit of engaging with the public. This is a collaborative, interdisciplinary project which will allow us to enhance and demonstrate the capabilities of innovative text mining tools we design to allow the automatic extraction of information from OCR-ed text. This paper presents the intended aims of the project; how this was achieved; an analysis of the findings; an interactive map of the results and a browser game of smells and disease.

Data and visualisation

As Roy Porter famously remarked that "todays history comes deodorised", sensory history is a relatively new historical approach. Historians rarely provide us an opportunity to hear, taste or smell the past. Medical historians have incorporated some aspects of sensory history into their research and explored the past belief that bad smells were causes of disease. However, there is very little research carried out covering this period.

Furthermore, smell has a great influence over how we perceive places and contributes to the construction of a place's identity (Quercia et al., 2015). During the 19th century the paranoia surrounding smells associated with poor hygiene heightened in many European cities (Reinarz, 2014). The Great Stink of 1858 resulted in the discussion of moving Parliament outside London for example. Despite the rise of germ theory (Pasteur and Koch) in the 1880s, concerns with disease-causing miasma (smells) did not disappear entirely. The MOH reports are one of the richest available sources on local public health administration and patterns of disease.

We enriched the text-mining pipeline with Natural Language Processing (NLP), including lemmatisation and part-of-speech tagging. The first iteration of the project has a feature to identify the category of the smells found by using a mapping table to work out the most common smell types. This step complements the close reading analysis and enables us to scale up the amount of information extracted from the texts. Our next research plan is to work on automatic identification of smell terms based on their contextual features to discover new categories that escaped previous classifications. This will allow us to identify smell categories in a data-driven fashion.

As the data becomes more structured, they can be more readily overlaid with other maps and images such as Charles Booth's London Poverty Map and 19th century disease maps. Having multiple layers will enable us to run various comparisons and assess if there are any correlations between smells and diseases as well as links to the socio-economic identity of areas in London.

During the first phase of the project we created a smelly map based on the number of smell hits to visualise the first set of results.

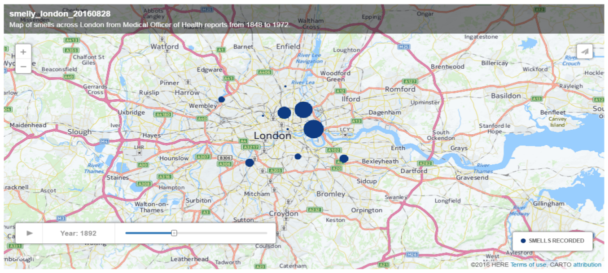
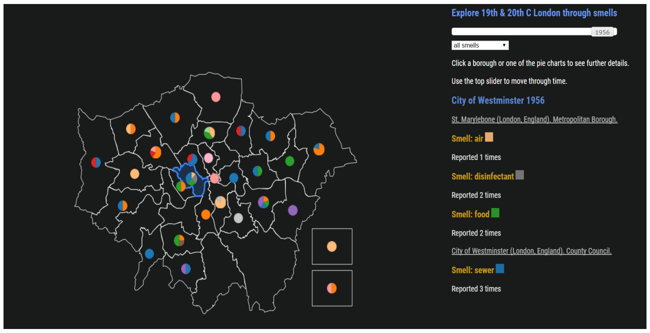


Figure 1. Smelly Map of London showing all smells

From the list of the existing London local authorities for the MOH reports we compiled, the geographic coordinates of present-day equivalents were extracted using an API. For the places that did not exist in the API, we manually added the geographic coordinates from Wikipedia. On the map each of the points marks the number of smells occurring at the centroid of each of the locations. We grouped the number of smells into sets of ten (e.g. 1- 10, 11 - 20) to avoid having giant points on the map for the places where there are almost 100 smells recorded. Finally, the map scrolls through the years. The data displayed in the mapping visualisation was obtained using text-mining via Python scripting. Python was the language of choice due to its high productivity rate and the fact that there are a large amount of third party libraries that offer highly useful functionality with just a few lines of code. For example, NLTK is a popular Python set of libraries that can achieve advanced NLP.

The next generation of map we produced during the second phase displays different smell categories that are colour-coded. The smell categories used for this map are Sewer; Waste-rubbish; Waste-excrement; Thames; Water; Food; Trade; Animal; Factory-fuel; Disinfectant; School; Air; Decomposition; Habitation; and Absence of smell. These categories were obtained through manual inspection of the data produced from searching for sentences containing smell-related words. In our codebase, we first analysed 5500 MOH London reports to find sentences that contained smell related words. Once the sentences were further analysed and categorised manually, the results were stored down into a local database by year, borough and a unique ID programmatically.


Explore 19th & 20th C London through smells

1956

all smells ▾

Click a borough or one of the pie charts to see further details.

Use the top slider to move through time.

City of Westminster 1956

St. Marylebone (London, England). Metropolitan Borough.

Smell: air 🟧

Reported 1 times

Smell: disinfectant ⬛

Reported 2 times

Smell: food 🟩

Reported 2 times

City of Westminster (London, England). County Council.

Smell: sewer 🟦

Reported 3 times

Figure 2. Smelly map of London showing smell categories

Computer programming can be used to perform tasks thousands of times faster than humans. In the Python code written to extract the data from the MOH reports, parallel processing was employed to speed up the running time of the program. Inside a computer there is a CPU which runs the tasks given by the program. Modern CPUs have multiple cores which allows the calculations to be run concurrently. In our project the CPU had four cores which allowed the running time of the program to be shortened by as much as three times. The next objective for the project is to scale up the size of the text-mining from 5,500 reports to over 70,000 reports covering the entire UK. In order to process such large datasets we are investigating the use of distributed computing resources such as Amazon Web Service (AWS). The code written for this project has been made open source under the MIT license along with documentation so that other programmers or researchers can use the codebase in their own text mining projects. The code has already been used in another project at Wellcome to investigate the idea of women's right to work during the 19th and 20th century London

Vision

The Smelly London project aims to bring together historical data with modern digitisation and visualisation to give us a unique, revealing and visceral glimpse into a London of the past and what it tells us about London today. Analysing the MOH reports tells the intimate narratives of the everyday experiences of 19th and 20th century Londoners through the 'smellscape'.

The Smelly London project provides a great opportunity to demonstrate how new knowledge and insights have risen from the use of powerful digital applications. This project will produce models that facilitate new kinds of humanities research. All outputs generated from the project will be open access and open source. Our data is available in a public repository on GitHub and other platforms.

Bibliography

Bynum, W. F. (1993) *Medicine and the Five Senses*, Cambridge; New York: Cambridge University Press.

Chave, S.. *Recalling the Medical Officer of Health: Writings by Sydney Chave*, London: King's Fund Publishing Office.

Classen, C, et al. (1994) *Aroma: The Cultural History of Smell*, London; New York: Routledge.

Cockayne, E. (2007). *Hubbub: Filth, Noise and Stench in England 1600 – 1700*, New Haven [Conn.]; London: Yale University Press.

Corbin, A. (1986) *The Foul and the Fragrant: odor and the French Social Imagination*, Leamington Spa: Berg.

Dobson, M. (1994). Malaria in England: A Geographical and Historical Perspective, *Parassitologia* 36 (1994): 35-60

Dobson, M. (1980)"Marsh fever"-The geography of malaria in England, *Journal of Historical Geography* 6(4) : 357-89.

Jenner, M. (2011) 'Follow your nose? Smell, smelling, and their histories', *The American Historical Review*, 116, 350

Quercia, D., Schifanella, R., Aiello, L. M., McLean, K. (2015). Smelly Maps: The Digital Life of Urban Smellscapes, *Proceeding of the 9th International AAAI Conference on Web and Social Media (ICWSM)*.

Reinarz, J. (2014) *Past Scents: Historical Perspectives on Smell*, Chicago: University of Illnois Press, 2014.

Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, et al. (2016) Text Mining the History of Medicine, *PLoS ONE* 11(1): e0144717. doi:10.1371/journal.pone.014471

Beyond Coocurrence: Network Visualization in the Civil War Governors of Kentucky Digital Documentary Edition

Patrick A. Lewis
patrick.lewis@ky.gov
Kentucky Historical Society, United States of America

Jeff Dycus
jeff.dycus@ky.gov
Kentucky Historical Society, United States of America

Anthony P. Curtis
tony.curtis@ky.gov
Kentucky Historical Society, United States of America

Whitney R. Smith
whitney.smith@ky.gov
Kentucky Historical Society, United States of America

Sara Carlsbad Brumfield
saracarl@gmail.com
Brumfield Labs, United States of America

Ben W. Brumfield
benwbrum@gmail.com
Brumfield Labs, United States of America

The Civil War Governors of Kentucky Digital Documentary Edition (CWG-K) is a freely-accessible online collection of historical documents associated with the chief executives of the state, 1860-1865. Yet CWG-K is about far more than the five governors. Uniquely positioned to view, reflect upon, and intervene in the lives of thousands of people experiencing the traumas of civil war, Kentucky's chief executives stood at the intersection of public and private life during an era that transformed the nation. Collectively, their papers provide a multi-faceted lens through which we can recover and understand more fully the lives of countless men, women, and children who have been heretofore both historically undocumented and archivally silenced.

After five years of active editorial work, CWG-K published the facsimiles and transcriptions of 10,000 documents online in 2016. Driven primarily by keyword searches and limited metadata faceting, *Early Access* is a digital evolution of the printed letterpress edition and its index. The content is interpretatively rich and suggestive and can be queried and sorted in a number of ways, but the experience is still ultimately linear. The ultimate goal of CWG-K, however, is to create a digital research environment within which a user can encounter the past multi-dimensionally through the documents and the powerful annotation network that links the documents together through the individuals, institutions, and places found in the texts.

CWG-K's true impact on scholarship, however, is through annotation. To the extent possible given the restrictions and biases of the historical record, CWG-K is identifying, researching, and linking together every person, place, and organization found in its documents. This web consisting of hundreds of thousands of networked nodes will dramatically expand the number of historical actors, show scholars new patterns and hidden relationships, and recognize the humanity and agency of historically marginalized people. The network of identified and annotated people, places, businesses, government agencies, and military units, will come as close as possible to a historical reconstruction of mid-nineteenth century society as it was lived and experienced in wartime Kentucky.

In this document-driven historical ecosystem, users can explore intuitively—moving seamlessly through seemingly disparate historical themes, events, and topics; breaking into the plane of social and geographic space to understand the deep patterns that underlay the issues raised in a text or set of texts; and moving forward and backwards through time to put this reconstructed historical world in motion and understand patterns of ebb and flow.

Phase II of the project (September 2016-September 2017) extends the edition into network visualization. Instead of reinventing the wheel for all the different functionality that an editorial annotation tool needs to have, we have taken the best existing open-source tools and integrated them to achieve our goals. TEI-XML files are exported from the transcription and control file tool DocTracker, published on the Omeka-powered Early Access website, and checked into a Github repository. Research assistants use Hypothes.is to identify entities within each document on the Early Access site. The custom-built MashBill tool queries Hypothes.is for those annotations, allowing research assistants to identify references to entities and build relationships between initially coocurring entities and document biographical research from external sources. MashBill then uses the annotations and identifications to update the TEI documents automatically with the appropriate persName/orgName/placeName references and republishing them to Github.

In the open source tool MashBill, CWG-K makes an important contribution to network analysis in digital humanities. With a few exceptions network analysis is still dominated by network construction based on cooccurence of entities within documents. These entities are most frequently created by automated named entity recognition performed on plaintext. CWG-K researcher's close reading of documents to extract entities allows for much accurate identification of entities. The use of Hypothes.is makes such human-powered entity recognition far more scalable than traditional manual tagging of TEI-XML. Rather than using pure cooccurence MashBill allows relationships to be defined by researchers consulting resources outside the documents of the edition. These relationships are then visualized with a D3.js visualization which is deeply linked to both the documents themselves as well as the articles written during the course of entity and relationship research on the people, places, and things. The open source MashBill tool may be reused for any TEI/Omeka project to reduce the effort and improve quality of entity identification.

Since the Civil War Governors of Kentucky Digital Documentary Edition is a project of the Kentucky Historical Society, it has always focused on public access. The development of MashBill and integration of the network visualization produced with that tool into the early access website will enable the public to discover the lives and stories of everyday people who interacted with the offices of the governors. Our synthesis of approaches and technologies provides an example other projects can benefit from, showing how to leverage open source tools and standards to efficiently identify and build network visualization in public digital editions.

Tackling Innovation Networks with Smart Data: A Case Study of the Liquid Crystal Institute at Kent State University

Hongshan Li
hli@kent.edu
Kent State University, United States of America

Marcia Zeng
mzeng@kent.edu
Kent State University, United States of America

Yin Zhang
yzhang4@kent.edu
Kent State University, United States of America

Xinyue Ye
xinyue.ye@gmail.com
Kent State University, United States of America

Tao Hu
taohu07@hotmail.com
Kent State University, United States of America

Numerous innovations and inventions have been made by human beings in the past few millennia, providing the most important driving force for economic, social, and cultural developments. However, for thousands of years, credits for most of these innovations and inventions have been given to individual genius-inventors in almost all civilizations. While the Chinese credited the Yellow Emperor for the invention of the compass and Cai Lun for papermaking, the British and the Americans honored James Watts and Wright Brothers for the invention of the steam engine and aviation respectively. This genius-inventor approach dominated the history of innovation for centuries, partly because it was difficult, if not impossible, for both professional and amateur historians to access or process information beyond a limited number of individuals involved in those creative activities.

The predominant position of the genius-inventor approach has been challenged when computers and other technologies invented in recent decades drastically increased human capability in acquiring and processing large amounts of data. By taking advantage of these new innovations, some scholars began to reexamine various major innovations by expanding their scope to cover the complex interactions between genius-inventors and their peers. With this new approach, F. C. Moon discovered that many of those innovations were actually the products of extensive interactions between the well-recognized geniuses and large social networks around them. While shedding new light on many of the innovations done in the industrial age or earlier, this new genius-centered network paradigm still pays the most attention to the central role played by the genius-inventors. This tendency limits the paradigm's ability to explain the numerous innovations and inventions made in the past century or so that did not involve any well-recognized individual geniuses. The invention of the liquid crystal display is one of those cases. Widely used in making TVs, computers, stationary and mobile phones, various control panels, billboards, watches, clocks, microwaves, and many other products that affect every aspect of our daily lives, the liquid crystal display was invented and improved by a large number of researchers who worked closely with their collaborators along the way. This absence of a single "genius inventor" in the development of the liquid crystal display and many other innovations calls for a new approach, one that will focus squarely on the large number of researchers who formed various networks that made those inventions possible.

This paper approaches the invention of liquid crystal display from a new perspective. Instead of focusing on genius-inventors or genius-inventor centered networks, this study examines all the researchers at Liquid Crystal Institute at Kent State University (LCIKSU) as nodes in an institutionalized network, and explores their interactions among themselves as well as with outside collaborators. LCIKSU is chosen not only because it is the largest research institute in the field that has made unparalleled and sustained contributions to the invention of liquid crystal display, but also because it has kept comprehensive institutional records since its inception, which makes it relatively easy to trace and analyze the growth and adaptation of this unique network of researchers.

The examination of LCIKSU is based on extensive data drawn from various sources. The research group, whose members come from the disciplines of Library and Information Science, History, and Geography, has been given the access to all the available institutional records, including the LCIKSU annual reports, grant and patent applications, conference and exchange information, etc. All living directors of LCIKSU, incumbent as well as retired, have given long interviews, some meeting with the group multiple times. In addition, a large amount of data has been collected from various other sources such as the number of LCIKSU researchers' publications from Web of Science, their grant awards from the National Science Foundation website, and their patent awards from ProQuest. With the help of various available data processing tools, usable information has been extracted from all these historical records and turned into smart data, which is then used to accurately measure the contributions made by LCIKSU to the invention of the liquid crystal display, and to decipher the secret of its success as a complex network of researchers.

The careful study of the LCIKSU history reveals that the success achieved in the invention of liquid crystal display in the past fifty years has depended on the establishment and maintenance of a dynamic and productive network of researchers who collaborated closely with each other both within and beyond their own subfields or disciplines. Instead of looking for a genius inventor or becoming one themselves, the LCIKSU directors made every effort to recruit collaborating scientists to cover various sub-fields related with liquid crystal research. Some of the scientists, between a handful to a dozen over time, emerged as primary nodes in the network, producing the most paper publications, winning the most grants, and receiving the most patents. Usually leaders in their own subfields, these primary nodes not only worked with each other in conducting cutting-edge research, but also built their own sub-networks through various means and extended collaboration with other scientists throughout the university system, across the nation, and around the world. In order to better understand and illustrate the success of the LCIKSU researchers, the smart data collected on their papers published, grants received, and patents awarded is processed using various network analysis tools so that the degree and frequency of their collaboration can be measured and analyzed. The resulting assortativity, average clustering coefficient, degree centrality, closeness centrality, and between centrality indicators fully support the assessments made by individual researchers that the concentration of a large number of scientists, especially the primary nodes, at LCIKSU and the close collaboration among them have made it possible for these networked scientists to produce more and better research results at greater speed. After all, it is this sustained high level of research that has made LCIKSU the lead inventor of the liquid crystal display.

Joseph Schumpeter, one of the best-known economists in the 20th century, accurately observed in the early 1940s that technological progress was increasingly becoming the business of teams of trained specialists. However, team innovators who have played a growing role in innovation and invention have not received adequate attention. This examination of LCIKSU not only puts the researchers and their networks in the spotlight of the invention of the liquid crystal display, but also introduces a new approach that sharply focuses on the networks of scientists and extensively uses smart data. With further development and refinement, this new network-centered and smart-data-based approach has the potential to help narrow the existing gap in the study of modern history of innovation.

Bibliography

Dunmur, D. and Kitzerow, H. (2016). The International Liquid Crystal Society 1990-2015. *Liquid Crystal Today*. 25 (2): 24-29.

Glänzel, W. and Schubert, A. S. (2004). "Analysing Scientific Networks Through Co-Authorship in Moed." Henk F. et al. eds. *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems.* Netherland: Kluwer Academic Publishers. 257-276.

Moon, F.C. (2014). *Social Network in the History of Innovation and Invention.* Springers Science.

Morgan, S., et al. eds. (2015). 50 Years of Innovation: Liquid Crystal Institute at Kent State University. Kent, OH: Kent State University College of Arts and Sciences.

Mowery, D., and Rosenberg, N. (1998). *Paths of Innovation: Technological Change in 20th-Century America.* New York: Cambridge University Press.

Schumpeter, J. (1942). *Capitalism, Socialism and Democracy.* New York: Harper & Row.

Teaching Empathy Through Virtual Reality

Amanda Licastro
amanda.licastro@gmail.com
Stevenson University, United States of America

In Philip K. Dick's Do Androids Dream of Electric Sheep the U.N. secretary proclaims, "[m]ankind needs more empathy" (1968). The poignancy of Dick's novel is its accurate expression of the social challenge of diminishing human empathy. The author offers empathy as the defining characteristic of humanity. As is often the case, science fiction foreshadows our future: longitudinal studies show decreasing rates of empathy in college students over the last three decades. If we believe that empathy is indeed a vital quality, then humanists are uniquely qualified to address this decline: extensive research suggests that empathy can be taught, specifically by reading fiction. Furthermore, preliminary trials indicate that virtual reality (VR) effectively evokes feelings of empathy in viewers. In both cases, the medium can provide the audience with access to situations outside of their everyday experience, offering a perspective into the lives of people unfamiliar to the reader/viewer. Take, for example, the work of documentary filmmaker Chris Milk that immerses the viewer in war torn villages in order to impact immigration policy (see "How virtual reality can create the ultimate empathy machine," 2016) or the content of the New York Times VR application which addresses a wide variety of social justice issues from all over the world. However, as critics such at Janet Murray rightfully argue, the impact of VR is dependent on the execution, which is still in development stages: "[t]he technical adventurism and grubby glamour of working in emerging technologies can make it hard to figure out what is good or bad from what is just new" ("Not a Film and Not an Empathy Machine," 2016). As the digital humanities have encountered with other emerging technologies - most notably data visualization techniques - these new forms need to be critiqued as they evolve (Drucker, 2012). Inviting students and educators to collaborate with industry professionals in the process of consuming, critiquing, and creating open access VR content creates the opportunity to design thoughtful immersive experiences that may address the decline in empathy in college age students. This presentation will explicate a study-in-progress devised to measure the pedagogical impact of VR content in combination with design thinking assignments used to combat desensitization and evoke empathy across the disciplines.

This research is supported with a case study of students in a series of linked courses at a small liberal arts college in Baltimore, MD. Students were exposed to VR content intended to increase their feelings of empathy for people who represent the "Other" in various ways, such as gender, race, ethnicity, and social class. This study was created through a cross-campus collaboration between faculty from the humanities, social sciences, and school of design alongside the theater director and librarians. Using empathy as the central question, each course integrated VR content and related readings into the curriculum. In each case, VR provided access to experiences not possible within the classroom space, for example an immersion into a refugee camp, a simulation of the human brain, and a documentary depicting gender bias across cultural contexts. The VR was scaffolded into each course in discipline-specific ways. For instance, the literature courses focused on readings that depict representations of virtual bodies in tandem with theory on posthumanism, particularly the work of Katherine Hayles and Donna Haraway. At the same time, the theater program produced The Nether by Jennifer Haley, which raises questions about the laws governing virtual spaces through depictions of pederasty and the murder of young children. Simultaneously, courses in psychology and human services integrated VR to discuss the impact of immersive content on social justice reform, and nursing courses looked at the application of VR for patient care and education. To varying degrees, this work was supplemented with readings on feminism, race theory, and disability studies in order to support discussions of "othering" with students. After analyzing the VR content in conjunction with the course materials, students were asked to design a VR experience intended to evoke empathy in the context of a discipline-specific audience. Additionally, members of a local VR company contributed as guest speakers and offered internships for interested students. Surveys were distributed at the beginning and end of the semester that prompted students to define, discuss, and debate empathy. At the end of each course students were interviewed to identify which methods of engagement increased their empathy toward people (in some cases characters) they felt were unlike themselves in significant ways.

As a part of this submission the syllabi and assignments will be shared. Ideally, the speaker will bring a VR headset and gaming laptop so participants can experience and consider how this emerging technology can evoke empathy by providing access to geographical, cultural, political, and biological content unfamiliar to the viewer. The goal is to receive audience feedback on the first stage of this study in order to improve and refine the methods before executing the plan on a larger scale. This study is IRB approved and student consent will be obtained for any student work that is presented.

Flexible Computing Services for Comparisons and Analyses of Classical Chinese Poetry

Chao-Lin Liu
chaolinliu@gmail.com
National Chengchi University, Taiwan

Introduction

As for many civilizations, poetry is an essential part of Chinese literature. Poetry has influenced the development of the literature and language of both classical and vernacular Chinese. Certain of the words that we use today can be tracked all the way back to the Shijing (詩經/shi1 jing1/ -- We show the pronunciations of Chinese characters in Hanyu Pinyin followed by their tones. Here, /shi1/ is for "詩" and /jin1/ is for "經".), c. 1046BC. Research on Chinese poetry is thus instrumental for understanding Chinese culture, and a lot of invaluable results have been accumulated over the past thousands of years from the study and analysis of Chinese poetry.

The availability of digital tools and resources enable researchers to compare and analyze the poetry from certain perspectives that were hard to achieve in the past. In many cases, we can verify the claims of previous researches with solid data, and, in others, we may enrich our understanding of the poetry.

The accessibility of increasingly larger datasets strengthens our research potential. In earlier stages of digital humanities, pioneers focused their work on Tang and Song poetry (it is beyond our capacity to list all previous research in this proposal, and we provide just two samples: Hu and Yu, 2001; and Lo et al, 1997) . Now, we can access digitized texts of poems that were published in the periods from 1046BC to modern days.

Software tools allow us to study the data from a wide variety of perspectives in an efficient way. Search engines and information retrieval techniques (Manning et al, 2008) help us extract relevant texts from a large dataset. Then, researchers can employ domain knowledge for advanced studies with the use of additional tools.

In this paper, we showcase research results that we achieved by handling the available data with existing tools in flexible ways. We collected nine representative corpora of Chinese poetry, one each for a major dynasty in Chinese history between 1046BC and 1644AD. We list the corpora in Table 1, where we assign an acronym to each corpus for ease of reference (See Notes, 1). We also show their Chinese names (**Collection**) and the periods of publication (**Time**). A collection for the Qing dynasty is unavailable yet because an editorial committee is still working toward the completion of this very challenging goal (Zhu, 1994). Excluding the punctuation marks that were added by the data providers, we have more than 16.5 million characters (see Notes, 2) in the corpora.

Acronym	Collection	Time	Acronym	Collection	Time
SJ	詩經	1046-476BC	**CV**	楚辭	475-221BC
HF	漢賦(文選)	202BC-420AD	**PT**	先秦漢魏晉南北朝詩	Before 589AD
CTP	全唐詩	618-907AD	**CSP**	全宋詩	960-1279AD
CSL	全宋詞	960-1279AD	**YSX**	元詩選	1271-1368AD
LCSJ	列朝詩集	1368-1644AD			

Table 1. The corpora of poetry of 1046BC-1644AD used in this study

By flexibly integrating and migrating tools to offer new functions, we can provide researchers with opportunities to investigate Chinese poetry from new perspectives. In the first example, we show a new way to compare the ways that poets use words in their poems. In the second, with our own tools, we can find shared collocations and patterns of poems in different corpora, and this capability allows us to study and compare the styles of individual poets and their dynasties.

	LSY	LB	DM	DF		LSY	LB	DM	DF
春風;秋風	14;2;16	72;26;98	18;11;29	19;30;49	春草;秋草	0;0;0	15;12;27	2;1;3	13; 5;18
春水;秋水	2; 3; 5	3;10;13	4; 5; 9	8;12;20	春色;秋色	0;1;1	9;11;20	3;6;9	20; 7;27
春月;秋月	0; 0; 0	0;40;40	0; 0; 0	0; 4; 4	春來;秋來	4;1;5	0; 3; 3	2;6;8	8; 6;14
春日;秋日	2; 2; 4	2; 1; 3	3; 1; 4	13; 5;18	春光;秋光	2;1;3	6; 0; 6	2;3;5	9; 1;10
春山;秋山	2; 0; 2	2; 6; 8	0; 4; 4	2; 5; 7	春天;秋天	1;0;1	2; 2; 4	0;0;0	5;11;16
春雨;秋雨	0; 2; 2	0; 2; 2	3; 3; 6	4; 4; 8	春江;秋江	0;1;1	1; 2; 3	0;2;2	6; 2; 8

Table 2. The frequencies of selected words used in poems of LSY, LB, DM, and MF

A Multi-Faceted Comparison

Jiang (2003) compared the usage of "wind" (風/feng1/) and "moon" (月/yue4/) in the poems of two of the most famous poets, Li Bai (李白/li3 bai2/) and Du Fu (杜甫/du4 fu4/), of the Tang Dynasty (which existed between 618 and 907AD) by comparing the contents of selected poems. Liu et al. (2015) listed the frequencies of frequent words that used "wind" and "moon" in Li's and Du's poems. The numerical comparison shows the differences of the poets in a vivid way.

The software tools can be designed so that we can inspect not just the original poems or the raw statistics about the original poems, but also more complex comparisons.

Table 2 lists the frequencies of frequent bigrams (again, see Notes, 2) that appeared in the poems of four poets, i.e., LSY (for 李商隱/li3 shang1 yin3/), LB (for Li Bai), DM (for 杜牧/du4 mu4/), and DF (for Du Fu) (Note: 李商隱/li3 shang1 yin3/ and杜牧/du4 mu4/ are two very famous poets of the Tang Dynasty)

These bigrams are special in that they are formed by concatenating either “春”/chun1/ or “秋”/qiu1/ with another character, and they represent something related to “spring” and “autumn”, respectively (when used individually, “春”/chun1/ or “秋”/qiu1/ typically represent “spring” and “autumn”, respectively– see Notes, 1). The numbers “14;2;16” in the row of “春風;秋風” and in the column for “LSY” indicate that we have 14 and 2 of LSY’s poems in which “春風” and “秋風” were used, respectively. “16” is the sum of 14 and 2.

The statistics in Table 2 shed light on the differences in word preferences among the poets. Note that the samples in Table 2 are limited, and that a close reading is necessary to reach any further interpretations. Despite these limitations, we still can explore comparisons from various perspectives. “春風” and “秋風” are the most common choices among all of the rows (They appeared 192 times, i.e., 16+98+29+49). In contrast, “春月” and “秋月” were not as popular (They appeared only 42 times, i.e., 40+2), and none of the poets used “春月”. In terms of personal preference, “春風” appeared in LB’s poems three times often than “秋風”. The results of LSY are similar to those fore LB, but DF seems to prefer “秋風” instead (The ratio for “春風”: “秋風” is 72:26 for Li Bai, 14:2 for Li Shang Yin, and 19:30 for Du Fu).

The entries that have “0”s can be linked to strong personal preferences. For instance, LB did not use “春雨” and “春來”, while he did use “秋雨” and “秋來”. DM is special in that he did not use “春天” or “秋天”.

We can provide different ways to compare the styles of poets, e.g., converting the frequencies in Table 2 to probabilities of seeing the same word in the poems. By building a vector space representation (Manning et al, 2008) for each poet, we can calculate a score of similarity for style as in many other researches.

Networking Names and Words

In addition to comparing the words of the famous poets, we may also attempt to compare the words and themes of the poems that were produced by friends. We can look up whether two poets were friends in professional databases like the China Biographical Database (CBDB) (Fuller, 2015). A database like the CBDB can also provide alternative names of poets so that we may algorithmically find friendships among poets by checking their writings (Liu et al, 2015). After identifying a group of poets who were friends, we can investigate whether the words, styles, and themes of their poems are related. A procedure such as which we used to produce statistics like those in Table 2 may be useful.

A poet may be influenced by another poet even though they were not personally acquainted. It is believed that poets of the same school of poetry (we use “school of poetry” to translate “詩派” /shi1 pai4/) share similar styles or words. Hence, information about the membership of a school of poetry provides a starting point for an investigation.

We may also search for poets who shared the same words and collocations in their poems as a clue for an indirect friendship. Given the millions of characters in our corpora, we need to have an efficient mechanism to identify poems that shared collocations and patterns (Liu and Luo, 2016), and using our own tools, we can precisely identify words that were shared by poems of different poets and of different dynasties (see Notes, 2).

The ability to identify the shared words between individual poets also automatically allows us to compare the patterns that are frequently shared between any two corpora. In Figure 1, two words are connected if they frequently co-occurred in poems. Part (a) shows the shared collocations in poems in the YSX and CTP, and (b) is for the shared collocations in poems of the LCSJ and CTP. The differences between (a) and (b) indicate that the highly shared collocations changed from dynasty to dynasty, i.e., from Tang to Yuan and from Tang to Ming. A collocation with a different word may suggest that the word contributes a different sense in the poems, e.g., “春風-桃花” and “春風-何處” in (b), and this can be verified by reading the poems that used these collocations. Sometimes, the links suggest replaceable words, for instance, both “千里” and “萬里” can go with “十年” in both (a) and (b). It should be noted that the collocations often carry information about the imagery of the poems.

Concluding Remarks

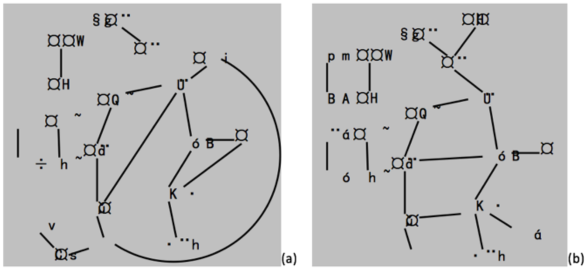
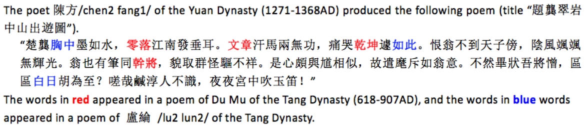


Figure 1. Frequently shared collocations between poems of two corpora (a) YSX and CTP (b) LCSJ and CTP

We briefly discussed how we studied two new research problems by flexible applications of our tools. The new tools provide new forms of data as in Table 2 and Figure 1 for interesting and useful research. We are working toward an in-depth understanding of Chinese words by studying when, who (Liu and Luo, 2016), and how the words (see Figure 2) and their collocations and patterns were used in Chinese poetry, and our tools will help domain experts study challenging and interesting problems about it (Liu, 2016). We also hope that the information and visualization that we have found and established for words can contribute to an interactive version of the complete Chinese lexicon (Cheng et al, 2016).



The poet 陳方/chen2 fang1/ of the Yuan Dynasty (1271-1368AD) produced the following poem (title "題龔翠岩中山出遊圖").

"楚龔胸中墨如水，零落江南發垂耳。文章汗馬兩無功，痛哭乾坤遽如此。恨翁不到天子傍，陰風颯颯無輝光。翁也有筆同幹將，貌取群怪驅不祥。是心頗與馗相似，故遣麾斥如翁意。不然畢狀吾將憎，區區白日胡為至？嗟哉鹹淳人不識，夜夜宮中吹玉笛！"

The words in **red** appeared in a poem of Du Mu of the Tang Dynasty (618-907AD), and the words in **blue** words appeared in a poem of 盧綸 /lu2 lun2/ of the Tang Dynasty.

Figure 2

Acknowledgments

This research was supported in part by the grant 104-2221-E-004-005-MY3 from the Ministry of Science and Technology of Taiwan, the grant USA-HAR-105-V02 from the Top University Strategic Alliance of Taiwan, and the Senior Fulbright Research Grant 2016-2017.

Notes

1. 詩經 /shi1 jing1/, 楚辭 /chu3 ci2/, 漢賦(文選) /han4 fu4 (wen2 xuan3)/, 先秦漢魏晉南北朝詩 /xian1 qin2 han4 wei4 jin4 nan2 bei3 chao2 shi1/, 全唐詩/quan2 tang2 shi1/, 全宋詩/quan2 song4 shi1/, 全宋詞/quan2 song4 ci2/, 元詩選/yuan2 shi1 xuan3/, 列朝詩集/lie4 chao2 shi1 ji2/

2. It is important to briefly mention the differences between Chinese characters and words for readers who are not familiar with the Chinese written language. Characters are the basic units for Chinese words. A Chinese word can be formed by one or more characters. For instance, "水"/shui3/ and "果"/guo3/ are two characters. They can be used individually to represent "water" and "results", respectively. A word consisting of *n* Chinese characters can be called an **n-gram** in linguistics, e.g., "水果" is a **bigram** that represents "fruit". While the majority of the words in vernacular Chinese are bigrams and trigrams, the proportion of unigrams in classical Chinese is very large.

Bibliography

Cheng, W.-H., Liu, C.-L., Chiu, W.-Y., and Hsu, C.-T. (2016). Phenomenology of emotion politics of color: Digital humanities research on the lyrical genealogy of 'White' in the poetry of the middle Tang dynasty (情感現象學與色彩政治學：中唐詩歌白色抒情系譜的數位人文研究), *Digital Humanities: Between Past, Present, and Future*, J. Hsiang (Ed.), 57–101, National Taiwan University Press.

Fuller, M. (2015). *The China Biographical Database User's Guide*, Harvard University. <http://projects.iq.harvard.edu/cbdb/home>

Hu, J. (胡俊峰) and Yu, S. (俞士汶). (2001). The computer aided research work of Chinese ancient poems, *ACTA Scientiarum Naturalium Universitatis Pekinensis*, 37(5):725–733.

Jiang, S.-Y., (蔣紹愚). 2003. "Moon" and "Wind" in Li Bai's and Du Fu's poems – Using computers for studying classical poems, *Proc. of the 1st Int'l Conf. on Literature and Information Technologies*. (in Chinese)

Liu, C.-L. (2016). Quantitative analyses of Chinese poetry of Tang and Song dynasties: Using changing colors and innovative terms as examples, *Proc. of the 2016 International Conference on Digital Humanities*, 260–262.

Liu, C.-L., and Luo, K.-F. (2016). Tracking words in Chinese poetry of Tang and Song dynasties with the China Biographical Database, *Proc. of the Workshop for Language Technology Resources and Tools for Digital Humanities*, The 26th International Conference on Computational Linguistics, 172–180

Liu, C.-L., Wang, H., Hsu, C.-T., Cheng, W.-H., and Chiu, W.-Y. (2015). Color aesthetics and social networks in complete Tang poems: Explorations and discoveries, *Proc. of the 29th Pacific Asia Conference on Language, Information and Computation*, 132–141.

Lo, F., (羅鳳珠), Li, Y., and Cao, W. (1997). A realization of computer aided support environment for studying classical Chinese poetry, *J. of Chinese Information Processing*, 1:27–36. (in Chinese).

Luo, Z. (罗竹风), ed., (1986). *Comprehensive Chinese Word Dictionary* (汉语大辞典), Shanghai Cishu Publisher (上海辞书出版社). (in Chinese) <http://hd.cnki.net/kxhd/>

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Owen, S. (2015) *The Poetry of Du Fu*. De Gruyter. <https://www.degruyter.com/view/product/246946>).

Zhu, Z.-J (朱则杰). (1994). Establishing the editorial board for the Complete Qing Poems (全清诗边簺筹备委员会成立), *Studies in Qing History* (清史研究), 0(3):96.

Character Distributions of Classical Chinese Literary Texts: Zipf's Law, Genres, and Epochs

Chao-Lin Liu
chaolin@nccu.edu.tw
National Chengchi University, Taiwan

Shuhua Zhang
shuhuazhang1@sina.com
Harvard University, United States of America

Yuanli Geng
geng99999@126.com
Harvard University, United States of America

Huei-ling Lai
hllai@nccu.edu.tw
National Chengchi University, Taiwan

Hongsu Wang
hongsuwang@fas.harvard.edu
Harvard University, United States of America

Introduction and Main Findings

Chinese characters are the basic units for Chinese words, and a Chinese word can include one, two, or more characters. Many characters can function as words in Chinese. For instance, “人文” represents “humanities”, and “人” and “文” are characters that carry their own meanings. Words that contain two, three, and more characters can be referred as *bigrams*, *trigrams* and other appropriate names for *n-grams*. Chinese does not separate words with spaces like English, so a reader must “segment” a character string into words to understand Chinese statements.

Mandarin Chinese has evolved over the past thousands of years (Yong and Peng 2008). Documents written in current vernacular Chinese contain a large number of bigrams and trigrams, while texts of classical Chinese (sometimes referred to as “literary Chinese”) contain many more unigrams. In the Academica Sinica Balanced Corpus (ASBC), unigrams contribute only 2.1% of word types, but constitute a 44.8% share of word tokens (“types” refers to distinct words in linguistics, e.g., “today is today” has three tokens and two types). In contrast, bigrams and trigrams together constitute 82% and 53% of the word types and tokens, respectively.

Zipf discovered that the relative frequencies of words are inversely proportional to their ranks in Chinese and English documents (Zipf, 1932) in his endeavor to establish a theory of *Principle of Least Effort* (Zipf, 1949). Many researchers have attempted to find the parameters in the Zipf-Mandelbrot distributions (see also, Piantadosi, 2015) for English and for Chinese (see Ha et al, 2003) many have explored extensions and applications of the law (see Altmann et al, 2002).

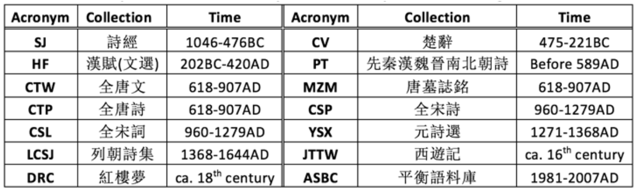
Previous research works for Chinese have focused mainly on fitting the Zipfian distributions to Chinese corpora. Some considered a distributed sample of corpora (Altmann et al, 2002), and others confined their analysis to a specific corpus (Deng et al, 2014). Some researchers have pondered on explanations for why the statistics of languages obey the Zipfian distributions (Powers, 1998, and Piantadosi, 2015).

We examined the Zipfian distributions of 14 collections of Chinese texts that were published from 1046BC (the year when the Zhou Dynasty (周朝) of China began) to 2007AD, and we found that the genres and epochs of the collections influence the distributions. The majority of our collections are poetic works written in classical Chinese. We also included official documents of the Tang Dynasty, novels of the Ming and Qing dynasties, and news articles of modern days.

The character distributions for the corpora of poems of 618-1644AD (the period stretching from the Tang dynastry to the Ming dynasty) exhibit strikingly similar Zipfian distributions. In contrast, the character distributions of the three genres of corpora that were all written in the Tang dynasty are distinguishable, although the characters frequently which are used in these documents are very similar. The word distribution of the ASBC differs significantly from other character distributions, indicating the importance of differentiating character- and word-based models of Chinese.

Corpora and Comparisons

For ease of references, we assigned an acronym for each of the 14 corpora, and show their names in Chinese (**Collection**) and periods of publication (**Time**) in Table 1.



Acronym	Collection	Time	Acronym	Collection	Time
SJ	詩經	1046-476BC	CV	楚辭	475-221BC
HF	漢賦(文選)	202BC-420AD	PT	先秦漢魏晉南北朝詩	Before 589AD
CTW	全唐文	618-907AD	MZM	唐墓誌銘	618-907AD
CTP	全唐詩	618-907AD	CSP	全宋詩	960-1279AD
CSL	全宋詞	960-1279AD	YSX	元詩選	1271-1368AD
LCSJ	列朝詩集	1368-1644AD	JTTW	西遊記	ca. 16th century
DRC	紅樓夢	ca. 18th century	ASBC	平衡語料庫	1981-2007AD

Table 1. The corpora used in this study include texts published during 1046BC-2007AD.

The corpora consist of representative literature that has been published since 1046BC. In particular, we have at least one collection for each of the major dynasties that existed before 1644AD. The majority of our collections are of poetic works (we consider SJ, CV, HF, PT, CTP, CSP, CSL, YSX, and LCSJ collections of poetic works) which fact lends itself to the study of the effects of genres on the character distributions. A collection of poetic works for the Qing Dynasty (which lasted from 1644 to 1912AD) is unavailable because an editorial committee is still working on its production (Zhu, 1994).

The corpora contain more than 42 million characters, excluding the punctuation marks that were added into the corpora by the data providers. When counting the characters, we also exclude characters that cannot be shown on ordinary computers. The frequencies of such rare and obsolete characters are not large, so ignoring them will not affect the statistical properties reported in this study.

Only the ASBC was segmented and the segmentation was verified by human experts. Hence, we can inspect its character and word distributions. The other corpora were written in classical Chinese and we do not have a reliable way for segmentation, so we will only analyze the character distributions.

We created charts that are based on the typical form of Zipf’s law:

$$\log\left(\frac{f(w)}{N}\right) = k - \alpha \log\left(r(w)\right)$$

where w, $f(w)$, and $r(w)$ denote a word, its frequency, and rank in a corpus, respectively. The rank of the most frequent word in a corpus is 1. N is the size of the corpus, and k and ✔are constants.

Observations and Discussions: Influences of Genres and Epochs

The generalizability of the Zipf’s law is the main reason that it has attracted the attention of many researchers. It can be applied to various natural distributions including those of part-of-speech of words (Wang et al, 2012), city sizes (Anderson and Ge, 2005), and corporal revenues (Chen et al, 2008).

Figure 1 shows the Zipfian curves when we consider the character distributions of all of the 14 corpora. We intentionally plot the curves in one chart, although this makes the individual curves undistinguishable. Although the curves are not linear, which is common as reported in the literature, the curves show a consistent trend, suggesting a common regularity that is shared by Chinese texts that were produced over the period of 3000 years.

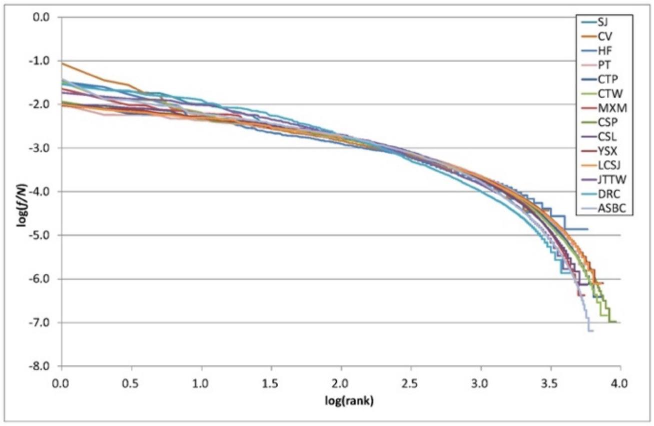


Figure 1. Zipfian curves of 14 corpora suggest a common trend. (Character distributions)

Instead of treating the 14 corpora as a single corpus to fit the resulting distribution for Zipf's law, we examined the curves to investigate possible factors that influenced the positions of the curves. In Figure 2, we show the curves of lyrics ("詞" /ci2/) and poems ("詩" /shi1/. The left halves of the curves overlap almost perfectly, which strongly indicates that the poetic works share very close statistical characteristics.

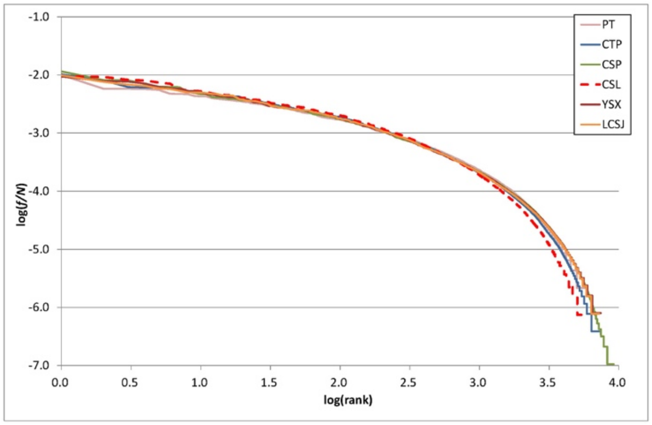


Figure 2. Zipfian curves of the corpora of lyrics (the red dashed curve) and poems (the rest) are strikingly similar. (Character distributions).

Table 2 lists ten most frequent characters found in each of the corpora and for which the curves are plotted in Figure 2.

Corpus	Ranks									
	1	2	3	4	5	6	7	8	9	10
PT	不	無	風	有	人	雲	之	何	日	我
CTP	不	人	山	無	風	一	日	雲	有	何
CSP	不	人	一	無	山	有	風	來	天	日
CSL	人	風	花	一	不	春	無	雲	來	天
YSX	不	人	山	風	一	雲	天	日	有	無
LCSJ	不	人	風	山	一	花	日	雲	有	無

Table 2. The most frequent ten characters in poems and lyrics remaining stable over time

The lists are very similar, and, out of the 60 characters in Table 2, there are only 16 distinct characters (some of which are homonyms, for which only one pronunciation is provided). In fact, we can compare the most frequent characters of any two corpora, e.g., the CTP and the CSP, to further investigate their similarity (Chen et al, 2012), and we found that the most frequent 1700 characters in the CTP and the CSP are the same characters.

Not all of the corpora of poetic works have similar curves. We added the curves for the SJ, CV, and HF in Figure 3, and it is evident that these new curves do not overlap with those in Figure 2 very well. The poetic works in the SJ, CV, and HF were produced very much earlier than those listed in Figure 2.

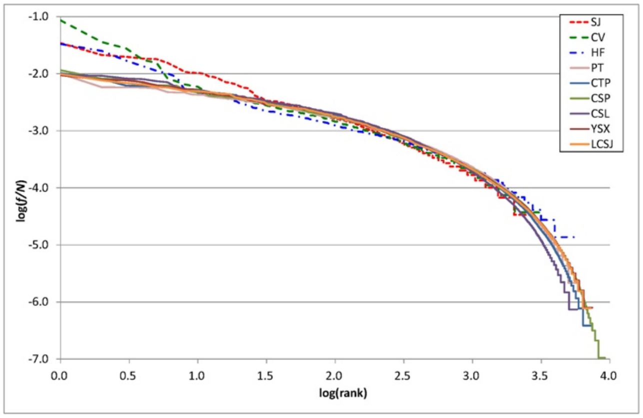


Figure 3. Curves for ancient poetic works (SJ, CV, and HF) do not coincide with those of later poems.

While the time of the production of the corpora affects the Zipfian curves, the curves for corpora that were produced in the same dynasty may not be the same. The CTW, MZM, and CTP are three different types of works that were all produced in the Tang Dynasty. We compared the most frequent characters shared by the CTP and CTW, and found that the sets of their most frequent 2000 characters differ only in three characters. Despite such an extreme overlap, their curves in Figure 4 suggest that genre affects the character distributions.

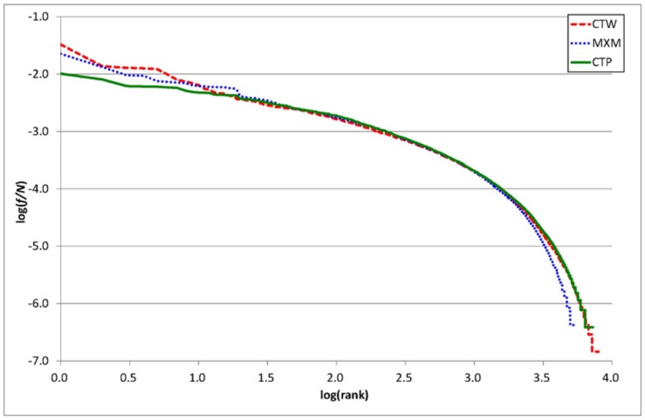


Figure 4. Curves of corpora that belong to the same dynasty but of different genres deviate from each other.

Given the above observations, one may have expected that the curves for the novels that were published in the 16th and 18th centuries, i.e., the JTTW and DRC, will deviate from the curves for the earlier poems, as the curves in Figure 5 show.

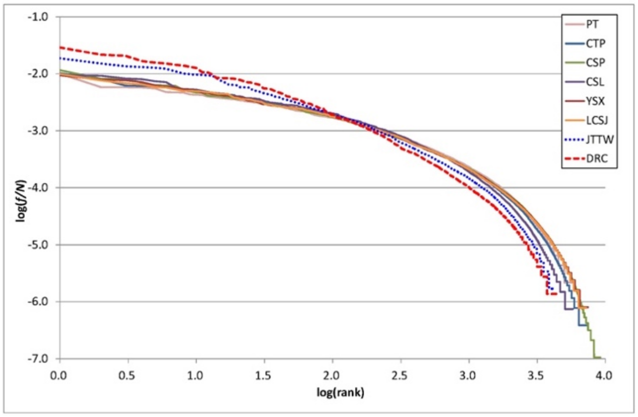


Figure 5. Curves of corpora that contain novels of 16th and 18th century (JTTW & DRC, respectively) deviate from the curves in Figure 2.

Character vs. Word distributions

We considered the character distribution when we analyzed the contents of the ASBC in Figure 1, where we found that the character distributions of the vernacular and classical Chinese texts show a reasonable common trend. The ASBC contains documents that were written in vernacular Chinese, so we must also analyze its word distribution, and Figure 6 shows the curves for both the character and word distributions.

A chart like that of Figure 6 can mislead one to infer that Chinese texts do not conform to Zipf's law. It is well accepted that the number of Chinese characters is limited, although there is no consensus about the exact number of characters. In contrast, there is virtually no limit on the number of legal Chinese n-grams. As a result, the sharp downturn of the character distribution and the intersection of the two curves in Figure 6 are expected, and this can be observed in languages other than Chinese in some special settings (Montemurro, 2001). We should examine the Zipfian curves on the same basis, e.g., character or word, while considering cultural factors that may influence actual language usage.

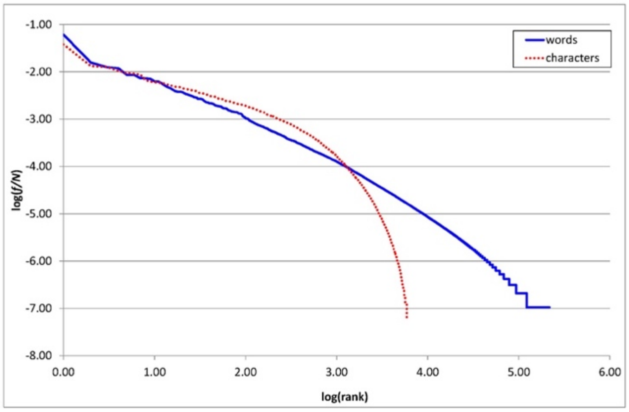


Figure 6. Word and character distributions for the ASBC differ significantly.

Concluding Remarks

We have judged the similarity between the Ziphian curves based on the visual closeness, though we can quantify the degree of similarity when desired (Hu and Kuo, 2005). Researchers have noticed the deviations of Zipfian curves at the high- and low-frequency ends (Hu and Kuo, 2005, Rousseau and Zhang, 1992) and tried to find density functions that fit the data. The statistics at the high-frequency ends of the curves are evidently more reliable. We focused on the deviations at the high-frequency ends of the curves, and discussed how the deviations in these regions may relate to the genres
and epochs of the corpora, employing the lists of most frequent characters of the corpora as extra supports.

Acknowledgments

This research was supported in part by the grant 104-2221-E-004-005-MY3 from the Ministry of Science and Technology of Taiwan, the grant USA-HAR-105-V02 from the Top University Strategic Alliance of Taiwan, and the Senior Fulbright Research Grant 2016-2017.

Bibliography

Altmann, G., Best, K.-H., Hřebíček, L., Köhler, R., Kromer, V., Rottmann, O., Schulz, A., Wimmer, G., and Ziegler, A. (Eds.) (2002) *Glottometrics* 5, RAM-Verlag.

Anderson, G. and Ge, Y. (2005). The size distribution of Chinese cities, *Regional Science and Urban Economics*, 35:756–776.

Chen, Q., Guo, J. , and Liu, Y. (2012). A statistical study on Chinese word and character usage in literatures from the Tang dynasty to the present, *Journal of Quantitative Linguistics*, 19(3):232–248.

Chen, Q., Zhang, J, and Wang, Y. (2008). The Zipf's law in the revenue of top 500 Chinese companies, *Proc. of the Fourth Int'l Conf. on Wireless Communications, Networking and Mobile Computing.*

Deng, W., Allahverdyan, A.E., Li, B., and Wang, Q.A. (2014). Rank-frequency relation for Chinese characters, *The European Physical Journal* B, 87, article 47.

Ha, L.Q., Sicilia-Garcia, E.I., Ming, J., and Smith, F.J. (2003). Extension of Zipf's law to word and character n-grams for English and Chinese, *Int'l J. of Computational Linguistics and Chinese Language Processing*, 8(1):77–102

Hu, C.-K. and Kuo, W.-C.. (2005). Universality and scaling in the statistical data of literary works, *POLA Forever: Festschrift in Honor of Professor William S.-Y. Wang on His 70th Birthday*, Dah-an Ho and Ovid J. L. Tzeng (Eds.), 115–139

Montemurro, M.A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics, *Physica* A, 300:567–578.

Piantadosi, S.T.. (2015). Zipf's word frequency law in natural language: A critical review and future directions, *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Powers, D.M.W. (1998). Applications and Explanations of Zipf's Law. *Proceedings of the Workshop on New Methods in Language Processing and Computational Natural Language Learning*, 151–160.

Rousseau, R. and Zhang, Q. (1992). Zipf's data on the frequency of Chinese words Revised, *Scientometrics*, 24(2):201–220.

Wang, D., Zhu, D., and Su, Z. (2012). Lotka phenomenon in the words' syntactic distribution complexity, *Scientometrics*, 90:483–498.

Yong, H. and Peng, J. (2008). *Chinese Lexicography*, Oxford University Press.

Zipf, G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*: *An Introduction of Human Ecology*, Addison-Wesley Press

Zhu, Z. (朱则杰). (1994). Establishing the editorial board for the Complete Qing Poems (全清诗边篹筹备委员会成立), *Studies in Qing History* (清史研究), 0(3):96. (in Chinese)

Lifting knowledge from the Medieval Age: CIDOC-CRM for the Nurcara Project

Matteo Lorenzini
matteo.lorenzini@oeaw.ac.at
Austrian Academy of Science, Austria

Luca Sanna
lucasanna@uniss.it
Università degli Studi di Sassari, Italy

Introduction

The structured repositories in cultural heritage have become the most used infrastructure for knowledge management towards different kinds of systems and platforms, ensuring a complete interoperability and reachability of data. Thanks to the semantic web paradigm, we are able to manage and enrich our data using formalisms and data standards: examples include digital libraries and digital archives, as well as SPARQL endpoints. However, the fragmentation of data produced by different kinds of mapping methodologies and different representations of the knowledge to be managed leads to some discrepancies between domains and results obtained during data retrieval.

A typical example is the study of an inscription, which will be addressed by linguists with regards to language; by philologists with regards to its text; by historians as a primary source; by archaeologists as material testimony of events and by conservationists as a piece of matter to be preserved and restored. In that scenario, semantic interoperability and standardization are two fundamental elements as they guarantee the circulation of knowledge inside a shared environment.

This paper aims to present the methodology followed in the Nurcara Project concerning resource integration and knowledge management. Nurcara consists of a dataset (almost 300 records) composed of textual Latin documents from the Medieval Age that provide historical context for the area of Monteleone Rocca Doria in Sardinia (Sassari, Italy) between the 11th and 15th centuries. Starting from an SQL database, our solution focuses on the development of a semantic framework solution able to both automatically map the relational dataset in CIDOC-CRM ontology on-the-fly and aggregate the knowledge from different repositories or endpoints. This method aims to semantically enrich Nurcara's dataset with external resources from the same domain, such as SPARQL-endpoint and Linked Data resources.

Nurcara project

Nurcara started from the following directive: to make accessible to the community the huge amount of published and unpublished medieval documents about Monteleone Rocca Doria that until now have been only accessible from the researchers.

Thanks to a collaboration between the Spanish Ministry of Cultural Heritage and the University of Sassari, it has been possible to start archival research of historical documents from the National Archive of Cagliari, the Historical Archive of Alghero and the General Archive of the Crown of Aragon in Spain. During the research almost 300 medieval documents in Latin, Catalan, Castilian and Sardinian were cataloged and stored in a MySQL database. The documents dated from the 11th to 15th centuries and related to Monteleone Rocca Doria (SS), which made them useful for defining the socio-economic context of the area during the medieval age.

Interoperability and semantic enrichment

The interoperability and semantic enrichment of the data represents one of the most important milestone of the project. It is crucial to publish the dataset in an accessible format with Semantic Web technology such as SPARQL or Linked (Open) Data.

In order to reach our goal, we started with the conceptualization of the entities from the relational database and the definition of the semantic schema. CIDOC-CRM has been chosen as the main reference model.

CIDOC-CRM is an ontology created in order to offer "definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation". The CIDOC-CRM model is a semantically rich model used to conceptualize the cultural heritage domain composed of 86 classes and 138 properties. The purpose of CIDOC-CRM is to provide a common definition for heterogeneous forms of information, and to enable their integration despite possible semantic and structural incompatibilities.

In order to translate data stored in the Nurcara relational database to the CIDOC-CRM ontology (expressed in RDF-Resource Description Framework) we have used the D2RQ tool that will provide an automatic mapping between the relational database and the CIDOC mapping schema.

D2RQ is a popular mapping platform for publishing relational data as a virtual RDF graph. It enables legacy relational databases to be exposed on the Web, according to the principles of Linked Data, and to be included in the Semantic Web. D2RQ exposes relational databases as SPARQL endpoints. It translates SPARQL queries posed on a virtual RDF graph to SQL queries posed to the underlying relational database.

Mapping Solution

The mapping process started with the definition of the main entities useful for the definition of the conceptual model directly from Nurcara's database:

- Document (E84 Information Carrier)
- Author (E39 Actor)
- Issue place (E53 place)
- Issue date (E50 Date)
- Type (E55 Type)
- Title (E35 title)

We chose these entities as a starting point for the mapping activities because from a conceptual point of view, they cover the minimum knowledge representation useful for the description of the documents from Nurcara's database. Furthermore, with the aim of defining a more suitable (event-oriented) semantic model with respect to CIDOC-CRM ontology, we defined some "abstract" entities that are outside the structure of Nurcara's database, but are present as classes in CIDOC ontology:

- Dataset (E73 Information Object)
- Event (E5 Event)
- Activity (E7 Activity)

Here, the use of the abstract classes as an intermediate layer allowed us to guarantee a more detailed and coherent conceptualization of the proposed model with respect to CIDOC-CRM ontology. Then, thanks to definition of the "Dataset" as E73 Information Object, it is possible to extend the proposed model with CRMdig (developed as a compatible extension of ISO21127) application profile from CIDOC-CRM ontology, which allows us to encode metadata about the steps and methods of production of digitization products.

The defined conceptual schema has been directly implemented as a mapping file (using D2RQ syntax) in D2RQ in .ttl format, in order to ensure the live mapping process between Nurcara dataset and CIDOC-CRM ontology.

In order to enrich and extend the dataset, we have also considered in the mapping.ttl schema other Linked Data end-points such as DbPedia or Geonames, which are specified by prefix and reachable via a proper SPARQL query directly from D2RQ SPARQL end-point.

Data accessibility

As previously stated, after the mapping process, the data are expressed and exposed as RDF graph based on CIDOC-CRM structure.

Hyperlinking

The D2R server supports hyperlink navigation by providing links on the RDF and XHTML levels. Any RDF triple whose object is a dereferenceable URI can be seen as a hyperlink. This is how resources published by the D2R Server are interlinked with other databases and external RDF documents. To aid discovery of related resources, D2R Server includes an rdfs:seeAlso triple with every resource description that points to an RDF document containing links to other resources produced by the same ClassMap (In our case, DbPedia or Geonames). If resources are identified with external URIs, then an additional rdfs:seeAlso link points to a local RDF/XML document that contains everything the database knows about the resource. By dereferencing the external URI and by following the rdf:seeAlso link, RDF browsers can retrieve both authoritative and non-authoritative information about the resource.

Search

The D2R Server allows the users to query non-RDF databases using the SPARQL query language over the SPARQL protocol. Queries are executed against a virtual RDF graph representing the complete database. Query results can be retrieved in the SPARQL query result as XML or in SPARQL/JSON serialization.

Conclusion

In spite of progress in the area of RDF storage, a large quantity of data is still stored in non-semantic repositories or relational databases.

Nevertheless, especially in the domain of the humanities, we are seeing growth in the use of semantic platforms for the management of digital data as a common method of knowledgement management. The CIDOC-CRM ontology is one of the most used in digital humanities. In this proposal we want to show how a specific dataset as Nurcara can be successfully integrated under the much more generic CIDOC-CRM structure, and how the knowledge can be easily enriched thanks to LOD integration. In order to reach the goal, the scalability and the full customization offered by the D2RQ server has played a crucial role.

The future development of this project concentrates on the integration of the Description Logic algorithm and reasoning system with the purpose of increasing resource discovery with respect to the domain utilized by the Nurcara project.

Bibliography

Bizer, C., and Seaborne, A. (2004) "D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs." Paper presented at the meeting of the ISWC2004 (posters), 2004.

Jannaschk, K., Rathje, C. A., Thalheim, B., and Förster, F. (2011) "A Generic Database Schema for CIDOC-CRM Data Management.." Paper presented at the meeting of the ADBIS (2).

Lourdi, I., Papatheodorou, C., and Doerr, M. (2009) "Semantic Integration of Collection Description: Combining CIDOC/CRM and Dublin Core Collections Application Profile." D-Lib Magazine 15 , no. 7/8.

Soddu, A. (2013) "Incastellamento in Sardegna:L'esempio di Monteleone." Castra Sardiniae. Quaderni Vol 1, ed. Lulu.com.

Mapping concepts and authors from 20th century Portuguese magazines of ideas and culture

Joana Malta
joanavmalta@gmail.com
Seminário Livre de História das Ideias
Universidade NOVA de Lisbon, Portugal

Pedro Lisboa
plisboa@gmail.com
Seminário Livre de História das Ideias
Universidade NOVA de Lisbon, Portugal

Magazines of ideas and culture constitute a very unique and exceptional set of sources, whose richness comes from an ample participation of intellectuals and artists, including some of the most influential individuals of their time. The fact that Fernando Pessoa first published most of his work in these periodicals clearly illustrates their importance (Andrade, 2003). These authors' prolonged participation, over issues published during long periods of time, allows us to capture and understand the movement of ideas and the diversity of their voices, in a continuous process of debate and update.

The periodical press played a very strong historical role in France (e.g., Charle, 2004; Pluet-Despatin et al., 2002), Italy (Frigessi, 1979), the United States (Tebbel and Zuckerman, 1991), and Brazil (Luca, 2011; Martins, 2012), to mention a few examples. The history of magazines in Brazil is particularly interesting to us, due to the very direct relationship between Brazilian and Portuguese intellectuals, who often wrote in the same publications (Guimarães et al., 2013).

In Portugal, all prominent intellectual movements used the periodical press, and specifically magazines, as a place of reunion, where they shared, discussed, and made their ideas public. In the first half of the 20th century the diversity of published magazines was as ample as the diversity of artistic, cultural, intellectual, and political movements. Magazines were used not only to reach other intellectuals but also a wider audience, with a commitment to establish a strong and informed public opinion, in search of what many contemporary authors believed to be "a universal rationality present in every person" (Andrade, 1999: 31, translation ours). The press was, among other things, a territory of political struggle and intervention.

For almost two decades now, the research group Free Seminar of History of Ideas (Seminário Livre de História das Ideias, n.d.) has been working on building a comprehensive and extensive database of some of the most important 20th century Portuguese magazines of ideas and culture. The project Magazines of Ideas and Culture (Seminário Livre de História das Ideias, Revistas de Ideias e Cultura, n.d.) has taken a multidisciplinary approach from the very beginning, encompassing knowledge from fields such as history of ideas, library science, and information science. Its aim was to build a relational database containing exhaustive information on authorship, quoted authors and works, subjects, concepts, and geographical names, for all articles contained in these publications.

One of the main methodological aspects of the project was the segmentation of the traditional "key-words" field in two: subjects and concepts. The key-word was deemed a too insufficient and imprecise instrument for obtaining a comprehensive understanding of the underlining ideas and ideological frameworks contained in these magazines.

To avoid the use of divergent criteria in establishing what could be considered a concept, a conceptual thesaurus, for use by researchers, was discussed and developed by the team beforehand. The first step in this process was to reach a consensual definition of the two relevant operative fields; in very broad terms, a concept was considered abstract, whereas a subject was deemed the materialization of an idea (e.g., Koselleck, 2004; Skinner, 2005 and Castro, 1996: 11-21, for a somewhat different theoretical approach). For example, in an article containing the concept "war," acceptable accompanying subjects – limited by actual article contents, of course – are any and all specific armed conflicts, as well as all the thoughts and ideas that can be produced on "war." We seek to determine concepts from a comprehensive, rather than explanatory, perspective. This means that, in any given article, all structuring concepts mobilized by the author, in accordance with the conceptual economy underlying the discourse, are relevant and thus collected.

Utilizing a conceptual framework also makes it possible to record the use of a concept without a verbatim textual mention of the word (Lisboa, 2015: 133-145). The need for a critical understanding of the texts made the use of tools for text recognition (OCR) inadequate. As such, all the input information is collected and validated by researchers.

The definition of the conceptual thesaurus followed principles of parsimony, consistency, and clarity, avoiding redundancies and repetitions, while at the same time favouring common criteria and language comprehensibility for all users:

> *"While making no claim to absolute denotation or universality, and agreeing with the fundamentally subjective use of concepts, we believe that the use of a large number of concepts is in fact ubiquitous, in parallel with the difference between language and speech. However, notwithstanding the differences between language and the uses of language, we must be able to achieve some level of systematization, as found in most dictionaries and similar works of reference*."

The resulting conceptual network constitutes one of the strongest virtues of the project, allowing for a clear delimitation of the general and specific terms that structured thought as a representation of the world, a phenomenon which can be found in magazines of the same political, cultural, or artistic movement. The set of reciprocal concepts in the conceptual map for each movement will allow a clear understanding of the fundamental ideas and thoughts that motivated their authors.

It should be noted that our project does not aim to classify information, but rather to provide a comprehensive reading of the sources. Our output is meant to be the result of a critical analysis of the source and an interpretation of its contents, grounded in the fields of History of Ideas and Conceptual History, and not a mere blind or neutral collection of terms and names. The database is not intended to replace the source, but instead to aid navigation through its complexity, providing meaning and structure to the included corpus of discourses, and highlighting the underlying programmatic and doctrinarian aspect of these magazines.

Our presentation proposal focuses on the magazine *A Águia*, a *monthly magazine of literature, art, science, philosophy, and social critique*, as stated in the subtitle, published between 1910 and 1932. This was the main organ of the republican Renascença Portuguesa, one of the most noteworthy intellectual movements of 20th century Portuguese history. For its 205 issues, published over the five series that make up the magazine's complete collection, 1,903 articles were indexed, and the collected data includes 403 single authors, 4,299 quoted names, and 1,033 concepts, among other descriptors.

We will apply network analysis to our data (Wasserman and Faust, 1994; Scott, 2000; Carrington, Scott and Wasserman, 2005), and intend to focus on the most important authors, i.e., those that have written the most articles for *A Águia*. Connecting individual article authors with the authors quoted in their texts should provide a visual map of intellectual networks, highlighting influences between authors and their ideas, thoughts, and intellectual background. Using information on quoted names will bring to light the shared and individual references of these authors, providing a visual network of the most influential names featured in their participation in the magazine.

We also intend to trace conceptual networks for a few selected authors. By observing the interaction of concepts used by a given author during a long period of time, an illustration of the course and evolution of their thoughts and ideas will hopefully manifest. In addition, observing the connections between concepts used by different authors (which concepts are used simultaneously more often, for example) can highlight profound or subtle differences in their intellectual background. Finally, confronting the two stages of our analysis – quoted authors and concepts – will provide a clear and vivid image of each author's views and discourse.

We believe that quantitative analysis of qualitative information can be a very powerful instrument if there is a strong consistency in the information being analysed. It can bring to light new relationships between authors, conceptual frameworks, or even conceptual transitions within a single author. Shifting from a narrative of intellectual dialogue to a visual movement of thought and ideas can be a powerful tool for obtaining an original and comprehensive image of the structures of thought underlying the intellectual production of a given period, through a particularly significant medium, i.e., magazines.

The tool developed by our research group "is an invitation for the reader to remember that knowledge is, foremost, an exercise of imagination, since the universe of questionable content and the scope of viable answers are both meticulously expanded." (Andrade, n.d.)

Bibliography

Andrade, L. C. de (2003). "Introdução: quatro notas breves." In Andrade, L. C. (ed), *Revistas, Ideias e Doutrinas. Leituras do Pensamento Contemporâneo*. Lisbon: Livros Horizonte, pp. 11-18.

Andrade, L. C. de (1999). "O Substantivo «intelectuais»." *Cadernos de Cultura*, 2: 23-41.

Andrade, L. C. de (n.d.). Revistas de Ideias e Cultura, http://ric.slhi.pt/A_Aguia/um_voo_singular_e_longo (accessed on 30th October 2016).

Carrington, P. J., Scott, J. and Wasserman, S. (eds) (2005). *Models and Methods in Social Network Analysis*. Cambridge: Cambridge University Press.

Castro, Z. O. de (1996). “Da história das ideias à história das ideias políticas.” *Cultura: Revista de História e Teoria das Ideias*, VIII (2): 11-21.

Charle, C. (2004). *Le Siècle de la Presse (1830-1939)*. Paris: Éditions du Seuil.

Frigessi, D. (ed) (1979). *La Cultura Italiana del '900 Attraverso le Riviste*. Turin: Giulio Einaudi.

Guimarães, L., Andrade, L. C. de and Castro, Z. O. de (2013). *Atlântida. A invenção da comunidade luso-brasileira*. Rio de Janeiro: Contracapa.

Koselleck, R. (2004). *Futures Past: On the Semantics of Historical Time*. New York: Columbia University Press.

Lisboa, P. (2015). “Edição electrónica de revistas históricas. O caso de A Águia.” In Rollo, M. F. and Amaro, A. R. (eds), *República e Republicanismo*. Coimbra: Caleidoscópio, pp. 133-145.

Luca, T. R. de (2011). *Leituras, projetos e (re)vista(s) do Brasil (1916-1944)*. São Paulo: UNESP.

Martins, A. L. (ed) (2012). *História da Imprensa no Brasil*. São Paulo: Contexto.

Pluet-Despatin, J., Leymarie, M. and Mollier, J.-Y. (eds) (2002). *La Belle Époque des Revues*. Paris: IMEC.

Scott, J. (2000). *Social Network Analysis: a handbook*. London: Sage.

Seminário Livre de História das Ideias, http://www.slhi.pt/ (accessed on 30th October 2016).

Seminário Livre de História das Ideias, Revistas de Ideias e Cultura, http://ric.slhi.pt/ (accessed on 30th October 2016).

Skinner, Q. (2005). *Visões da Política: sobre os métodos históricos*. Algés: Difel.

Tebbel, J. and Zuckerman, M. E. (1991). *The Magazine in America, 1741-1990*. New York: Oxford University Press.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: methods and applications*. Cambridge: Cambridge University Press.

Topic Patterns in an Academic Literary Journal: The Case Of *Teksty Drugie*

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research
Polish Academy of Sciences, Poland

Maciej Eder
maciejeder@gmail.com
Institute of Polish Language
Polish Academy of Sciences, Poland

Modelling Literary Scholarship

The availability of digitised full-text resources, as well as bibliographical data in standard database format, has recently opened a new chapter in the sociology of literature by revaluating empirical approaches and data-driven scholarship. The road to this “empirical turn” in literary scholarship has been paved by such scholars as Franco Moretti (2005, 2013) and Matthew Jockers (2013), who showed how empirical data like bibliographical records, annotations, title words, genre categorization, etc., may help in generating new knowledge about literary periods. This approach gathered its momentum as other works exploring the possibility of using such data to answer particular research questions emerged. Due to the shortage of space we will name just a few that have the most influence on this paper, dividing them into three research strands. Firstly, the use of bibliographical data for statistical inferences on literary processes, e.g Bode’s (2012) rereading of Australian literary history through the data from AustLit (Australian Literary Bibliography). Secondly, the study of author co-occurrences and mutual references, e.g. visualising literary circles on the basis of such data by Long and So (2013a, 2013b). Thirdly, the application of topic modelling to uncover pertinent issues in literary scholarship, e.g. Goldstone and Underwood’s analyses of the evolution of American literary scholarship on the example of PMLA (2012) and seven major literary journals (2014). In combining those approaches into a macroanalytical study of *Teksty Drugie*, we also adopted the rationale introduced by the 40th anniversary internet edition of *Signs*, a literary journal dedicated to feminist criticism.

Aim

The aim of this study is to apply macroanalytical methods to trace the chronology of transformations of Polish literary studies using the example of *Teksty Drugie.* We hypothesise that the collection of papers published in a leading academic journal on literary scholarship can serve as a reliable approximation to chronological changes and/or breaks in Polish literary theory at the turn of the 20th century. We will first trace the topics present in the journal and then analyse them in diachronic perspective. We will focus on the influence of extra-textual events and phenomena on literary scholarship.

We believe that 25 years is a sufficient timespan to observe linguistic differences which are not caused by regular language change. Other projects conducted by the authors of this paper show that a language change (in Polish) typically spans many decades rather than a mere 25 years (e.g. Eder & Górski, 2016). Furthermore, we deal with conventionalised language of scholarship, so the use of certain terms often relates to a given paradigm rather than to a language in general. Nevertheless, we are aware of possible changes of meaning of keywords while interpreting topic models.

Material

Teksty Drugie is a Polish literary journal dedicated to literary scholarship. It has been published since 1990 by the Institute of Literary Research of the Polish Academy of Sciences. It focuses on literary theory, criticism and cultural studies, while also publishing articles by authors from neighbouring disciplines (philosophy, sociology, anthropology). The journal publishes monographic issues dedicated to particular topics or approaches within literary and cultural studies. All those features make it a good example for exploring the vicissitudes of Polish literary scholarship.

The corpus consists of the entire collection of papers published in *Teksty Drugie* (excluding letters, surveys, notes, etc.) in the years 1990–2014 (2,553 texts, 11,310,638 words). The material covering the years 1990–1998 was digitised, OCR-ed, and then manually edited, in order to exclude running heads, editorial comments, and so forth. Obviously, some textual noise – e.g. a certain number of misspelled characters – could not be neutralised. The material from 1999 onwards was digitally-born, but even though a small number of textual issues might have occurred. We believe, however, that distant reading techniques are resistant to small amounts of systematic noise (Eder, 2013).

Given the nature of Polish, which is highly inflected, lemmatization was necessary for a reliable processing of texts. The corpus has been lemmatised with LEM 1.0. (Literary Exploration Machine) developed by CLARIN-PL (see: Piasecki, Walkowiak, Maryl 2017).

Method

To scrutinise the formulated hypothesis, we applied one of the methods of information retrieval that recently attracts a good share of attention in Digital Humanities circles, namely topic modelling in its classical variant known as Latent Dirichlet Allocation (LDA). The method, introduced by Blei (2012), allows for finding co-occurring cohorts of words that presumably reveal (latent) semantic relations.

The experiments were performed using a tailored script in the R programming language, supplemented by the package 'stylo' (Eder et al., 2016) for text pre-processing, and the package 'mallet' (McCallum, 2002) for the actual LDA analysis. A bimodal network of the relations between topics were produced using the software Gephi (Bastian et al., 2009).

Topic modelling relies on the assumption that particular topics are defined by words co-appearing in a given context. Hence, the definition of "context" is crucial to allow for any reliable observations. A few different solutions have been suggested (e.g. Blei, 2012; Jockers, 2013). In our approach, we did not split input texts into smaller samples, which was motivated by the fact that the vast majority of the studies published in Teksty Drugie are rather short.

Other parameters used in the study included: a stop word list containing 327 words (mostly function words, numerals, and very common adverbs), 100 topics extracted in 1,000 iterations, with the obvious caveat that this choice was arbitrary.

Results

A general overview of the obtained results shows a few interesting patterns. Firstly, we analysed and categorised the topics on the basis of their predominant words. The categories are as follows: literary theory (e.g. literature, fiction, text), poetics (e.g. verse, novel, short story, rhetoric) and methodological approaches (e.g. deconstruction, comparative literature, postcolonial studies, psychoanalysis); history of literature (e.g. romanticism, contemporary poets) and cross-cutting research themes (e.g. death, politics, literacy).

A thorough exploration of such models requires a topographical visualisation capable of showing the connections between various topics, which often share a key word (cf. Goldstone and Underwood, 2012). The network (Fig. 1) is too large to be adequately rendered in this paper (a higher resolution image of Figure 1 is available online), yet even without the knowledge about concrete topics presented, we may see (partly thanks to ForceAtlas2 layout, which highlighted this feature) that groups of topics in our corpus are concentrically distributed. This onion-like distribution allows us to distinguish between the central topics (i.e. those who appear in many different papers) and those who appear less often or sporadically and hence are not particularly well-connected with other topics. For instance, in the geometrical centre of the network we may find topics and words pertinent to literary scholarship: literature, literary, comparative literature, national literatures, Jewish studies, fiction, together with some names of contemporary authors. Outliers are also interesting, and could be assigned to 3 groups: (1) expressions in foreign languages, (2) particular research topics or discourses which introduce quite a hermetic language, not shared in other topics, (3) noise (e.g. word bits generated through some errors in OCR).

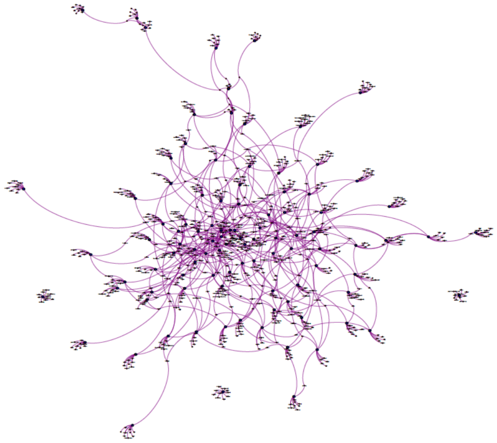


Figure 1. Relationships between topics in Teksty Drugie.

Yet it has to be noted that even the most accurate rendering of the topical distribution is still only a static snapshot insensitive to changes. In order to see the evolution of topics, we need to visualise them on a temporal axis. Due to a shortage of space we present here only a few examples, to show the application of our method. All dot plots are presented below with a trend line based on two period moving average.

Fig. 2 represents the gradual shift of interest from more literature-oriented approaches, to the cultural ones. Both red (topic 19: literature, literary, writer, work) and green (topic 5: literature, research, theory) seem to be dominating until approx. 2007, when the blue line (topic 49: culture, cultural, social) overtakes the green line for the first time. Three years later it becomes the dominant approach, marking the shift in the overall content of Teksty Drugie.

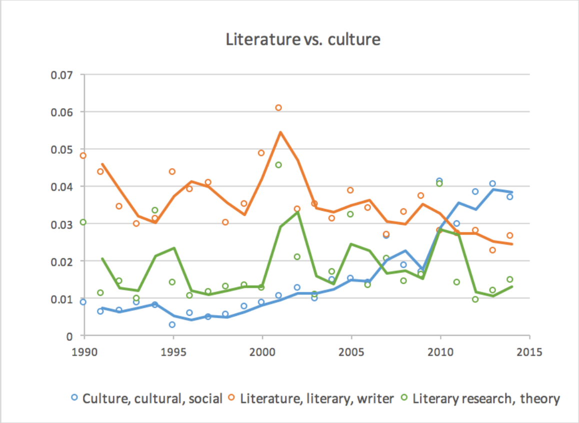


Fig. 2. A temporal distribution of three topics related to literature and culture.

Topic analysis allows us to not only trace the evolution of the journal itself but also to see how the real-world events shape the topics undertaken by literary scholars. Fig. 3 shows the influence of the political transformation in Poland on the content of Teksty Drugie. We see a similar pattern in trends of all topics presented: grey (topic 60: power, society, state, fight, war, law), red (topic 36: political, communism, Polish People's Republic), blue (topic 7: Polish, Pole, national), yellow (topic 94: censorship, exile, novel, positivism, country, London, political). All of them are quite important in the early 1990s and the interest gradually fades until the end of this decade. The spikes around 2001/2002 are caused by the publication of monographic issues which make certain topic more dominant. E.g. Issue No.1-2/2000 was dedicated to socialist realism hence the spike of "communism-related" issue in that year.

This trend shows how political events (namely the transformation and forming of the new democracy) are dominating even the literary scholarship. It could be also the case that more politically charged issues (e.g. history of censorship in Poland) could have been published only after the fall of the communism, hence so many articles in that period.

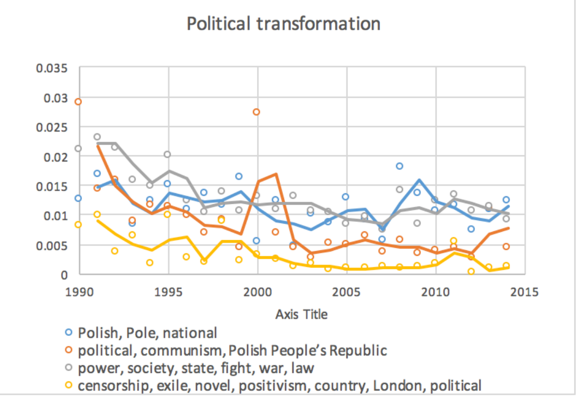


Fig. 3. Temporal shift of topics related to politics.

The last trend we would like to discuss is the emergence of the Holocaust studies in Teksty Drugie. As we can see in the Fig. 4, the red trend line (topic 59: Jew, Jewish, antisemitic) is visible on the fairly same level all through the 25 years, whereas the blue one (topic 18: testimony, Holocaust) is virtually non-existent until 2001.

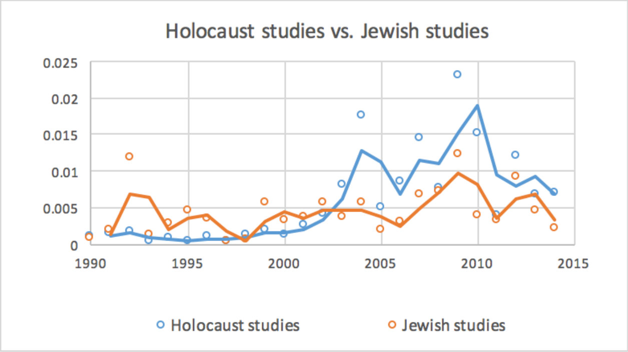


Fig. 4. Temporal distribution of topics related to Jewish studies and the Holocaust.

This sudden boom can be linked to the publishing of the Polish edition of Neighbors by Jan Gross (2000) and the investigation into the role of Polish civilians in the genocide perpetrated in the city of Jedwabne during the World War II. This case opened a long process of re-investigating the troubled Polish-Jewish past, which could be traced also in the issues of *Teksty Drugie*.

Conclusions

In this study we tried to show how extra-textual events influence the content of literary scholarship on the example of Holocaust studies and political transformation, which entailed the prevalence of topics related to politics, power, society, state, and communism in the early 1990s. In the subsequent studies we plan to compare the results of topic modelling with bibliographical data in order to check whether the dominance of a certain topic stems from the large number of scholars who pursue it, or if it instead depends on the fact that a small group of authors published more often than others.

Acknowledgements

The authors wish to thank Dr Tomasz Walkowiak for his extensive help with cleaning and lemmatizing the corpus.

Bibliography

Bastian, M., Heymann, S. and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International ICWSM Conference*. San Jose, pp. 361–62.

Bode, K. (2012). *Reading by Numbers: Recalibrating the Literary Field*. London & New York: Anthem Press.

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4): 77–84.

Eder, M. (2013). Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing,* 28(4): 603–14.

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal,* 8(1): 107–21.

Eder, M., Górski, R. (2016). Historical Linguistics' New Toys, or Stylometry Applied to the Study of Language Change. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 182-184.

Goldstone, A. and Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship?. *Journal of Digital Humanities*, 2(1).

Goldstone, A. and Underwood, T. (2014). The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3): 359–84.

Gross, J. T. (2000). *Sąsiedzi: Historia zagłady żydowskiego miasteczka.* Sejny: Fundacja Pogranicze.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press.

Long, H. and So, R. (2013a). Network science and literary history. *Leonardo*, 46(3): 274–274.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Moretti, F. (2013). Distant Reading. New York: Verso Books.

Piasecki, M., Walkowiak, T., Maryl, M. (2017). Literary Exploration Machine (LEM 1.0) - New Tool for Distant Readers of Polish Literature Collections. Paper accepted for presentation at ADHO Digital Humanities conference at McGill Universiy, Montreal

So, R. and Long, H. (2013b). Network analysis and the sociology of Modernism. *Boundary* 2, 40(2): 147–82.

Mapping the Meshwork of the Independent Media Arts

Lindsay Kistler Mattock
lindsay-mattock@uiowa.edu
University of Iowa, United States of America

During the late-1960s and early 1970s, independent media artists imagined a network of organizations that would support the production, distribution, exhibition, preservation, and study of film and video not only in the known centers of activity (New York City and San Francisco), but across all regions of the United States. By 1980 the Media Arts Center Movement had gained significant momentum leading to the establishment of the National Alliance of Media Arts Centers (NAMAC). These centers included national players, such as the Museum of Modern Art in New York City, along with regional organizations like the Rocky Mountain Film Center and Pacific Film Archives, and other metropolitan organizations like Pittsburgh Filmmakers (one of the oldest remaining active media arts centers). These organizations provided services for artists – funding and equipment rentals – and to the surrounding community – screenings, coursework, and study collections. NAMAC changed its nomenclature in the mid-1990s to the National Alliance of Media Arts and Culture, yet the legacy of the media arts center movement continues today. Mapping the Independent Media Community (MIMC) is a project that seeks to illustrate the impact of the individuals and organizations that were part of this larger movement to support the development of independent media arts, not just in the United States, but across the globe.

MIMC is currently in its first phase of development, generously funded by a Major Project Grant from the University of Iowa. Partnering with Carnegie Museum of Art (CMOA) in Pittsburgh, Pennsylvania, the MIMC team at the University of Iowa has worked in conjunction with CMOA's Time-Based Media Project (funded by a grant from the A.W. Mellon Foundation) to digitize and provide access to records contained with the CMOA's Film Section Archives (spanning 1970-2002). The MIMC prototype is based on the data contained within the *Film and Video Makers' Travel Sheet* and *Film and Video Makers' Directory* published between 1973 and 1987. The *Travel Sheet*, a monthly publication, served as a social networking tool for media artists and media arts centers. At no cost, artists could publish booked tour dates along with a general sense of their travel schedule, while organizations posted the contact information for their internal programmers, indicating a willingness to host makers. Over the years, the *Travel Sheet* grew to include sections for new works available, film festival announcements, and other advertisements of general interest to film and video makers. The *Directory*, published in 1978 and 1979 included the information of all of the individuals and organizations subscribing to or listed in the *Travel Sheet*. From its humble beginnings as a single hand-typed 11x17 sheet of paper, the publication grew to include thousands of individuals and organizations from all corners of the globe. Using the data from the *Travel Sheet,* this first phase of the MIMC project serves as a prototype for a much larger and more robust application that will visualize data from a wide variety of sources in order to provide a more comprehensive understanding of the Media Arts Center Movement in the United States as well as global independent media production, distribution, exhibition, preservation, and study.

The MIMC database (built using the University of Sydney's Heurist Academic Knowledge System) currently holds only a few years of the *Travel Sheet* data. However, even this small subset of the data MIMC illustrates that the Media Arts Center Movement was indeed successful in building a wide-reaching network of organizations and individuals. The image below represents a sample of data from New York-based artists reported between 1973 and 1975 in the *Film and Video Makers Travel Sheet* (data from 26 artists residing in New York City and New York state).

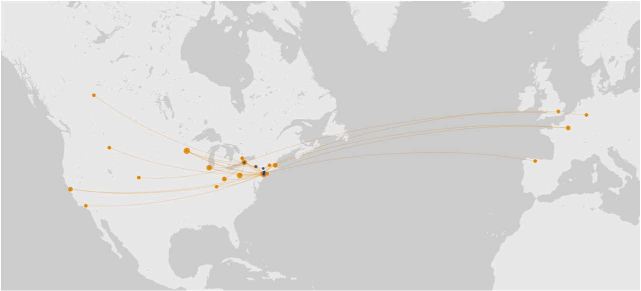


Figure 1: Selected event data from the *Film and Video Makers Travel Sheet* 1973-1975 (created with Palladio).

In the above image, the artists' home address is represented in blue while the event location is represented in orange (the event locations are sized according to the number of events, 99 total). This visualization demonstrates the connections between organizations and individuals emerging from the data in the *Travel Sheet.* Here we can see the beginnings of a network that extends beyond the United States and into Europe and Canada. As data entry and analysis continues, the data promises to demonstrate a global network of artists and media arts organizations as a sample from the 1979 *Film and Video Makers' Directory* illustrates, below.

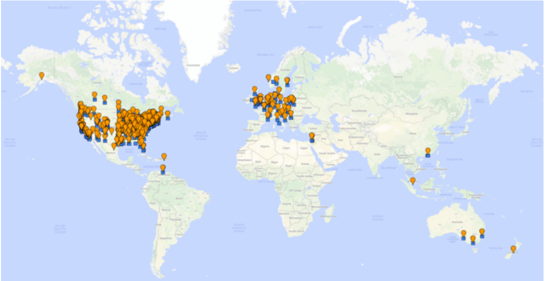


Figure 2: Organizations (square) and Individuals (teardrop) from the 1979 *Film and Video Makers Directory*

While the term network was the preferred terminology of the Media Arts Center Movement, this is an overly simplistic understanding that masks the complexity of the MIMC data. Anthropologist Tim Ingold argues that the lines of a network are connectors, static points joining two nodes, that presume an absolute connection that does not fully illustrate the complexity of the relationship. In contrast, Ingold offers the concept of the meshwork, suggesting that these lines are not connectors, but the "lines of becoming" from Deluze and Guattari's rhizome (Ingold, 2013: 132). These lines do not meet – the nodes in the network map instead represent knots and entanglements in the meshwork, places where lines emerge and diverge rather than connecting in absolutes. The meshwork, like the rhizome, affords multiple narratives and entry points; it offers no hierarchy or structure. MIMC seeks to offer this alternative mapping of the meshwork – the entanglements between artists, distributors, museums, governmental bodies, local communities, and countless other actors in the fabric of the independent media arts.

Unlike other digital humanities approaches to cinema and media history, like Jeffrey Klenotic's Mapping Movies, MIMC provides access not only to a representation of the history of the independent media arts, but access to the archives that hold these traces as well. The MIMC data model includes the provenance for each discrete data point in the database, linking individual records back to the primary source material from which it was derived. These connections will afford opportunities to link directly to the digital archives that hold the digital surrogates as these records, allowing MIMC to serve as an extension of these archival collections as a digital finding aid of sorts. In this way, MIMC is enfolded in the archives; the project does not derive-from but is entangled-with the archival organizations and other sites that hold the history that MIMC seeks to represent, continuing to build the meshwork that entangles the various organizations and individuals represented in the visualizations. In locating and documenting these archival traces, MIMC provides an understanding not only of the historical unfolding of the independent media arts and Media Arts Center Movement, but of the archivalization of this history as well – the project itself becoming further enmeshed and entangled in the very history that it seeks to uncover.

This brief paper will introduce the MIMC project and discuss the development of the MIMC application as well as the potential impact of the project as a Public Digital Humanities resource for scholars and for the archives that collect and provide access to the primary source materials from which the MIMC data is derived. In addition to the database and visualizations, by preserving the source information for each record and linking to the digitized archival records (or archival finding aids), MIMC links the archive of the independent media arts that is distributed across archives, personal collections, the active and inactive organizations that are part of this vast meshwork. While there is still much work to be done, MIMC promises to provide widespread access to historical data that can be reused and re-imagined beyond the initial bounds of the project.

Bibliography

Ingold, T.(2013). Making: Anthropology, Archaeology, Art and Architecture, New York: Routledge

Chatbot Based Content Discovery: Faulknerbot in the Archive

Aaron Mauro
mauro@psu.edu
Pennsylvania State Digital Humanities Lab
United States of America

Introduction

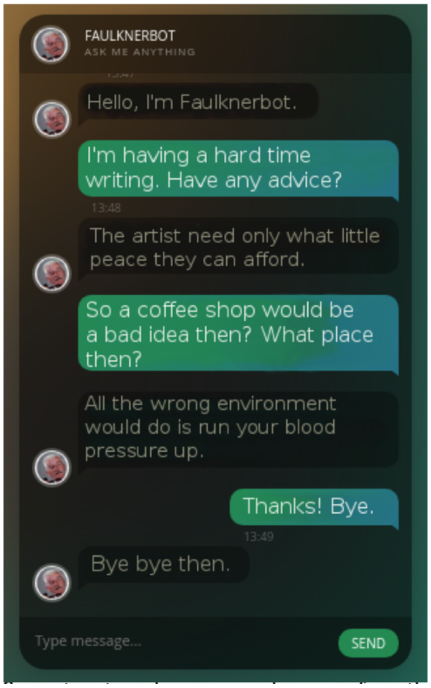
In March of 2016, the failure of Microsoft's proto-type chatbot, Tay, was not just a technological failure. It was a disciplinary failure. It was a failure of an indus-try leader to adopt a critical perspective when build-ing systems in a complex cultural and social environment. Tay, which stands for "thinking about you," was the name given to an artificial intelligence chatbot for Twitter that was quickly corrupted by users and began spewing racist, sexist, and homophobic slurs. Pundits quickly leapt to conclusions about the political beliefs of internet users, but these same pundits failed to un-derstand that this hacking of Tay was in fact a critique of chatbots in the real world. Users of Twitter were exposing a fundamental error made by the Microsoft development team. Because the system learned directly from user input without editorial control or content awareness, Tay was quickly trained to repeat slurs by users eager to embarrass Microsoft.

This moment in technological development makes for an interesting anecdote, but it also represents the moment that chatbots entered the public conscious-ness and became nothing less than the future direction of a unified interface for the whole of the web. Of course, chatbots captured imaginations in the 90s as well. Systems like Cleverbot, Jabberwacky, and Splotchy were fascinating to play with, but they had no real application. Today, text based AI has been identified as the the successor to keyword search. No longer will we plug in keywords into Google, comb through lists of text, and depend on search engine optimization (SEO) to deliver the best content. Search will be around for a long time, but in the near future much more content will be delivered through text based messenger services and voice controlled systems. We've seen the early stages of this change in products like Amazon's Alexa, Apple's Siri, Google Now, and Microsoft's Cortana. There are now bots embedded within common platforms like Slack, Skype, and Facebook Messager. We are now approaching a world that Apple envisioned in 1987 with a mockup system called the "Knowledge Navigator" that sought to give users an interactive and intelligent tool to access, synthesize, present, and share information seamlessly.

Humanities in the Loop

We are likely decades away from a true "knowledge navigator," but the second generation of these chatbots are now in development. The company that developed Siri for Apple is now in the final stages of development on a system called Viv (Matney). Viv is the first viable company to produce a unified interface for text and speech based AI assistants. Facebook is testing project M within its messenger app to allow users to issue commands, access services, and make purchases through text input (Hempel). The remarkable thing about M is that Facebook has built a system with "humans in the loop." This means that when a service is accessed, perhaps by purchasing movie tickets, a human will fine tune the AI generated results for each transaction. There is currently an understanding within the machine learning community that human assisted training of these systems produces more accurate results but will also train more robust systems go-ing forward (Biewald, Bridgwater). The current need for human in the loop systems means that we are at a crucial moment for humanists to lend their experience and critical abilities to the development and training of AI systems. In the field of machine learning, training a system to answer humanities based problems will show how these systems succeed or fail, but they will also demonstrate the value of the humanities in a digital world. If the purpose of the humanities is to better understand what it is to be human, training AI to answer philosophical, historical, or cultural questions will help us understand our experiences as we become more accustomed to intelligent systems in our lives. Grappling with AI, whether it is in a mundane consumer exchange or in matters of grave ethical importance, is rapidly becoming a practical problem in our lives.

With humanists in the loop, we will better under-stand the social and cultural contexts in which these systems appear and avoid the regrettable failure of systems like Tay in the future. We are currently on the cusp of a revolution in the applicability of natural language understanding, artificial intelligence, and conversation based interfaces design. These technologies will have ranging consequences socially, culturally, and economically in the coming decade, but these technologies are also deeply connected to the social and cultural contexts in which they appear. My goal is to train machines to be humanists. It is the literary critic's ability to close read complex philosophical, historical, and artistic meaning that these systems lack. It is the ability of the historian to contextualize political and technological change within the breadth of human progress. It is the dramatist's ability to understand performance and dialogue that will animate our conversations with computers. The digital humanities are well situated to make the most of NLP techniques and find culturally significant training sets.



FAULKNERBOT

ASK ME ANYTHING

Hello, I'm Faulknerbot.

I'm having a hard time writing. Have any advice?

13:48

The artist need only what little peace they can afford.

So a coffee shop would be a bad idea then? What place then?

All the wrong environment would do is run your blood pressure up.

Thanks! Bye.

13:49

Bye bye then.

Figure 1: Faulknerbot interface with basic query and response

Method: Conversational Data Retrieval

Biographical and archival material has been used to train a system to allow a conversation with the famed American author William Faulkner. I will present a system trained with nearly all the interviews that Faulkner has given. Author interviews are an excellent training set because the questions asked by the interview anticipate user interests and model a conversational style of response. The interviews collected during Faulkner's visit to the University of Virginia were instrumental in building this tool. The applications for such a system are numerous. A conversation with Faulkner might benefit a creative writing student in the midst of writer's block. A chatbot offers a more inviting interface for a general public. Most importantly, Faulknerbot will represent a novel form of content discovery for student researchers. Once a user has developed a chat history worth exploring, Faulknerbot's responses link to original archival materials for research purposes.

Current systems have come a long way from the toy-like chatbots that populated the web in the late 90s. After a preprocessing stage using word2vec, which vectorizes the bag of words, this model uses Tensor Flow to generate two complementary neural networks that encodes and decodes inputs and responses. This model has only recently been made accessible to non-computer science researchers recently by Google open sourcing Tensor Flow. is not based on the retrieval based model using a rule based expressions, with a heuristic to determine intent and draw from a predefined response. This is not a simplistic tree model based on nested "if/then" statements. Instead this uses a generative model. This generative model uses sequence to sequence learning with neural networks. This model links words statistically to determine "flows" of meaning through a word vector. Geoff Hinton calls this a "thought vector." In other words, this is an end-to-end model that remains open. Rather than a retrieval method, which limits the scope of the conversation, this system dynamically learns and allows for a retention of what has been said. The generative model allows for this context based discussion without resort-ing to an enormous conversation log. In Tensor Flow, this operates on a Long Short Term Memory (LSTM) network. As I've said, the sequence to sequence model is based on two neural nets. One is an encoder, which encodes input data from the user. The decoder model determines the reply by generating the output, which need not echo the size of the vector. This thought vector generalizes input and links to a target response. This is not a "feed forward" neural net. It is a recurrent neural net that continually retrains on the training data, which is often the marker of a true "deep learning" system. This model makes no assumption about purpose or predetermined output. It simply reinforces relationships between thought vectors over time. There is a deeply emotional resonance that is carried through conversation. The blurring of lines between social media, search, and messaging will result in a seamless and unified interface for digital technology. Driven by the mobile space's demand for streamlined UI design, we will become more reliant on assistive technologies that can anticipate, learn, and adapt to user input.

Conclusion

It is important for the humanities to anticipate this new cultural space. When the Google autocomplete system was introduced to search, there were many cultural commentators decrying the loss of independent thought and the potential for entrenching damaging stereotypes (Postcolonial). The loss of critical awareness and even just the ability to spell. Technology that offends our sense of what it is to be essentially human is usually the next important media type. Chatting with machines tends to cross such lines. There are practical uses for remedial education and composition studies. A functioning Teaching Assistant Bot capable of answering questions about deadlines, assignments, and course policy would be welcome by most educators. Indeed, an AI TA has been developed recently, but it is unclear if this system can be trained on any course material or was custom built for this class (Maderer). Generalizing these systems is a difficult task, to be sure. The newly open sourced Tensor Flow machine learning library can answer questions derived from a training set of just over a million words. When we consider the limits of machine learning in intelligent assistants, scholarly communication through chat interfaces is certainly the next logical step. However, these systems require humans in the loop. They require thoughtful and critical reflection. They require an attention to depth and nuanced meaning. They require a humanist in the loop.

Bibliography

Alphabet. (2011) Google Code Archive. Web. <https://code.google.com/archive/p/word2vec/>.

Apple (2016) "Knowledge Navigator." 28 October. <https://www.youtube.com/watch?v=HGYFEI6uLy0>.

Bridgwater, A. (2016) "Machine Learning Needs A Human-In-The-Loop." 7 March. <http://www.forbes.com/sites/adrianbridgwater/2016/03/07/machine-learning-needs-a-human-in-the-loop/#2175c4ba6590>.

Biewald, L. "Why human-in-the-loop computing is the future of machine learning." 13 Nov. 2015.<http://www.computerworld.com/article/3004013/robotics/why-human-in-the-loop-computing-is-the-future-of-machine-learning.html>.

Carpenter, R. Cleverbot. 28 Oct. 2016. <http://www.cleverbot.com/>.

Carpenter, R. Jabberwacky. 28 Oct. 2016. <http://www.jabberwacky.com/>.

Railton, S. et al. (n.d.)Faulkner at Virginia. <http://faulkner.lib.virginia.edu/>.

Hempel, J. "Facebook Launches M, Its Bold Answer to Siri and Cortana." 26 Aug. 2015. <https://www.wired.com/2015/08/facebook-launches-m-new-kind-virtual-assistant/>.

Hunt, E. (2016). "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter." 24 March .Web. <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter>.

Maderer, J. (2016) "Artificial Intelligence Course Creates AI Teaching Assistant." 9 May 2016. <http://www.news.gatech.edu/2016/05/09/artificial-intelligence-course-creates-ai-teaching-assistant>.

Matney, L. (2016). "Siri-creator shows off first public demo of Viv, 'the intelligent interface for everything'" TechCrunch. 9 May.<https://techcrunch.com/2016/05/09/siri-creator-shows-off-first-public-demo-of-viv-the-intelligent-interface-for-everything/>.

Microsoft. (2016) "TayTweets." Twitter. 28 October. <https://twitter.com/tayandyou>.

Postcolonial Digital Humanities. (2013) "Google's Autocompletion: Algorithms, Stereotypes and Accountabil-ity." 19 November. <http://dhpoco.org/blog/2013/11/19/googles-autocompletion-algorithms-stereotypes-and-accountability/>.

Splotchy (2016). Algebra.com 28 Oct. <https://www.algebra.com/cgi-bin/chat.mpl>.

Credit Allocation with the Social Knowledge Timeline

Aaron Mauro
mauro@psu.edu
Pennsylvania State Digital Humanities Lab
United States of America

Introduction

Humanities scholarship is becoming increasingly collaborative, participatory, and public facing. As humanists take up digital tools to conduct and share research, larger teams are needed to complete ever more complex computational tasks. When blending these heterogeneous teams--which may include faculty, librarians, staff, undergraduates, graduate students, postdocs, and community contributors--humanists have an ethical responsibility to offer a fair and transparent accounting of research activities. Tracing the evolution of research contributions is necessary for a range of issues facing digital scholarship such as authorship allocation, promotion and tenure, and reports to funders. The allocation of credit and authorship is an increasingly thorny issue for teams with a range of possible roles and a variety of research outputs and media types. There is, however, a large amount of data being generated by these teams that is capable of describing and measuring the contributions made on a variety of platforms and by multiple team member and community partners.

Despite our inheritance of social and collaborative tools, many of these systems elide the nuance and process of humanities based research. Knowledge creation is not merely a function of how much code is produced. New knowledge is often the result of a key insight made by a team of students, staff, and faculty. These insights are generated in a complex and overlapping system of mentorship, service, teaching, learning, and authorship that are deeply dependent, social, and human. With a system that is possible of visualizing the history of a digital project over the course of years, which may see the ranks of team members change over time, primary investigators and project funders will be better able to address often thorny and ethically charged issues relating to student assessment, mentorship, authorship, promotion, and tenure. Credit, promotion, funding, and credentialing are more complex topics than ever, yet many individuals and institutions rely on simple, outdated structures to assess the value of insights made by networked teams.

Social Knowledge Creation

The Penn State Digital Humanities Lab (Penn State Behrend) in partnership with the Teaching and Learning with Technology group (Penn State University Park) has developed a prototype of an ongoing project entitled the Social Knowledge Timeline (sktimeline.net). By linking together popular collaboration tools, the SKTimeline stores, analyzes, and communicates user data in three distinct areas of social knowledge creation:

- **Collaboration Platforms:** Many scholars are turning to collaboration platforms like Slack, Yammer, and Basecamp to organize teams and foster communication within teams. These systems use an interface similar to a social media feed to pool project member input into a single narrative and eliminate the need for email. These systems help share documents and support conversations that may lead to drafting manuscripts on Google Drive and other services.

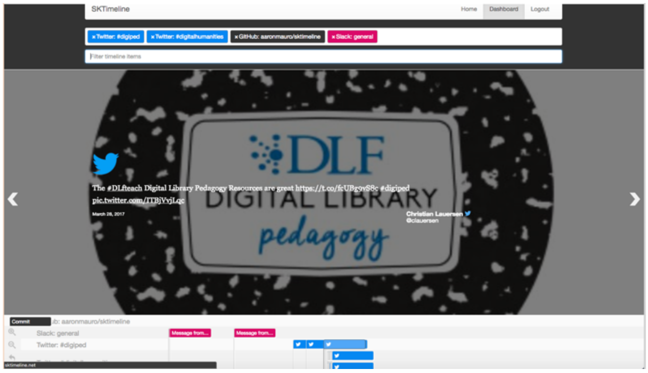
These platforms offer a rich, conversational natural language data set that describes how team members mentor and support each other over time.

- **Version Control Systems:** Github and Bitbucket are two of the most common version control platforms. These tools help facilitate large programming and encoding projects by allowing multiple coders to work simultaneously. When a team member "commits" code, a commit message describes the nature of the contribution as well as the date and time. This message will offer a highly granular view of coding projects as they unfold. Similarly, by including a feed from Google Drive's own version control system, document authorship may be traced with similar precision.
- **Social Media:** Platforms like Twitter, LinkedIn, and Facebook have proven to be fast paced and engaging areas for social and cultural exchange. Twitter has long been a particularly important site for digital humanists. The SKTimeline draws together multiple hashtags and user handles to frame preserve and contextualize this often ephemeral site of both popular and scholarly debate. Hashtags associated with digital projects, conferences, publications, and even course work can be analyzed and set in real time with other platforms.

Credit allocation in large teams is dependent on our ability to describe, quantify, and visualize our activities. By analyzing the rich natural language conversations generated by teams, the SKTimeline solves these ethical and institutional problems. The appearance of "Collaborators' Bill of Rights" for digital humanities projects in 2011 is symptom of a need for greater clarity in heterogeneous collaborative teams (Clement et al 2011). The Modern Language Association's "Guidelines for Evaluating Work in Digital Humanities and Digital Media" are similarly responding to appropriate credit allocation for researchers. There is a need for a more formalized and automated system of data collection and analysis for collaborative researchers across the university.

Machine Learning Contributor Taxonomies

The Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) is used to quantify and describe user contributions. Machine learning systems like Google's Cloud Platform is used to conduct language analysis, and translation, image recognition, sentiment analysis, and keyword extraction. Custom machine learning systems has also been layered on to these services using the Tensor Flow library to learn the project specific phrasing for contributions. Additional text analysis will be conducted using standard tools like the Natural Language Toolkit (NLTK) to link to TaDiRAH's defined contributions. This project will reshape authorship and credit allocation in the humanities and beyond, but it will also be a perfect test bed for an emerging set of artificial intelligence tools that are now finding common application throughout society. In this way, the SKTimeline is representative of a broader cultural trend toward AI systems in aiding research.



The #DLfteach Digital Library Pedagogy Resources are great https://t.co/fcUBg9vS8c #digiped pic.twitter.com/TTBjVvjLqc

March 28, 2017

Christian Lauersen
@clauersen

Figure 1. The Social Knowledge Timeline displaying Slack channels posts and Twitter hashtags chronologically. Images associated with posts are used for backgrounds on the timeline

Conclusion

Undergraduate course projects, ongoing faculty research with graduate researchers, digital humanities labs, and library based digital research projects are just some of the contexts this round of user testing will examine. The data collected on participating teams against interview and form based user surveys. This kind of socially oriented knowledge creation emerges from a community of practice that moves fluidly between curricular experiences and co-curricular research experiences often hosted in DH labs, libraries, and centers. The SKTimeline seeks to solve a critical problem within scholarly communication in a digital context. The SKTimeline offers a means to capture complex narratives that constitute the organic and nuanced unfolding of humanities research.

Bibliography

Alphabet. (n.d.) Google Cloud Platform. <https://cloud.google.com/products/machine-learning/>

Alphabet. (n.d.) Tensor Flow. <https://www.tensorflow.org/>.

Borek, L. (n.d.) TaDiRAH. <https://github.com/dhtaxonomy/TaDiRAH>.

Clement et al. Eds. (2011) Off the Tracks: Laying New Lines for Digital Humanities Scholars.

Media Commons Press, (2011). Web. <http://mcpress.media-commons.org/offthetracks/>. Committee on Information Technology (2012) "Guidelines for Evaluating Work in Digital Humanities and Digital Media." Modern Language Association. Web. <https://www.mla.org/About-Us/Governance/Committees/Committee-Listings/Professional-Issues/Committee-on-Information-Technology/Guidelines-for-Evaluating-Work-in-Digital-Humanities-and-Digital-Media>.

Di Pressi et al. (2015). "A Student Collaborators' Bill of Rights." UCLA Digital Humanities. Web.

<http://www.cdh.ucla.edu/news-events/a-student-collaborators-bill-of-rights/>.

Python Software Foundation. (n.d.) Python Language Reference, version 3.6. Available at <http://www.python.org>.

Ronacher, A. (n.d.) Flask. <http://flask.pocoo.org/>.

Research Center as Distant Publisher: Developing Non-Consumptive Compliant Open Data Worksets to Support New Modes of Inquiry

Robert H. McDonald
rhmcdona@indiana.edu
Indiana University, United States of America

Introduction

In the original Google Books Settlement Agreement in 2008 (Courant 2009), funds were to be set aside to create a research center that would enable researchers worldwide to accomplish data-mining and analysis on texts in the public domain and under copyright in a manner that was secure and compliant with appropriate U.S. copyright law. This did not happen, because the court rejected the agreement in 2011. Despite this, in 2011, the HTDL announced that Indiana University Bloomington and the University of Illinois at Urbana-Champaign would run the HTRC under a cooperative funding agreement with the HathiTrust Board of Governors and the University of Michigan. Since 2014, HTRC has made available as an active production service tools to analyze a set of out-of-copyright content equaling around 4.4 million volumes. In 2016, the HTRC plans to enable analysis of the entirety of the 15 million-volume corpus currently held by the HTDL, the largest digital academic library in North America.

HTRC and Non-Consumptive Research

The HTRC has developed a process to define and work within the concept of *non-consumptive* computational access to support the fair-use of the HTDL corpus as defined within the Google Books Settlement Agreement that was a part of the *Authors Guild et al. v. Google Inc* case.

Currently the HTRC defines the process for *non-consumptive* use of the HTDL corpus as:

> Research in which computational analysis is performed on one or more books, but not research in which a researcher reads or displays.

Operationally, from the perspective of the HTRC research cyberinfrastructure, the HTRC defines *non-consumptive* research as:

> That which requires that no action or set of actions on the part of users, either acting alone or in cooperation with other users over the duration of one or multiple sessions can result in sufficient information gathered from a collection of copyrighted works to reassemble pages from the collection.

This concept has been further refined in the course of the development of the HTRC Data Capsule (Zeng et al. 2014) for secure data analysis and the development of the HTRC Workset Ontology (Jett et al. 2016) and has been codified in the recently released HathiTrust Research Center Non-Consumptive Use Research Policy (HTRC, 2016).

HTRC as Publisher

During the course of work with scholars using the HTRC tools and services to create derivative non-consumptive data sets, the Center has often taken on a set of the roles traditionally played by publishers. These data sets are reviewed by members of the HTRC staff for compliance with non-consumptive use standards prior to release to the authors.

As part of this work, the HTRC has offered as a service the capability to publish these non-consumptive, compliant data sets using a DOI scheme (Downie; 2015). This service enables the creation of new derivatives (Downie; 2015) of published non-consumptive compliant data sets.

A second benefit of opening access to these data sets is the ability to replicate current experiments that have been developed using the HTDL corpus and the HTRC tool set. From this standpoint the HTRC functions as a *distant publisher* of non-consumptive compliant data sets in support of new models of research inquiry.

Distant Publishing as Concept

Prior to defining the concept of *distant publishing*, it is first instructive to understand *distant reading* within the context of digital humanities. *Distant reading* was first codified in 2000 by noted humanist and scholar Franco Moretti:

> Distant reading: where distance . . . is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, less is more. (Moretti 2000)

Moretti later expanded the concept in his 2013 monograph of the same name (Moretti 2013). Much like Moretti's definition that focuses on enabling a broader view of the text, the *distant publisher* enables a broader view of data sets through bringing to bear the current corpus of computational tools for large-scale textual data mining and analysis.

HTRC as a distant publisher is removed by at least one degree from the creator, and remains distinct from any standardized concept of publisher. Yet, data sets are published under the rubric of the HTRC, and these publications are freed from the constraints of copyright in this context due to their non-consumptive nature. Thus we define distant publishing as:

> Publication of a non-consumptive data set outside of any standardized publishing construct, removed by x degree from the original creator, openly available to the community of scholars for replication and available for re-use in support of the advancement of knowledge.

This definition is one that the HTRC aims to further refine in the coming years. We welcome broader thoughts on this concept from those working to preserve open research data and the software that makes that data accessible for use in scientific experimental replication and re-use for the long-term benefit of the scholarly community.

Distant Publishing Use Cases

Currently the HTRC is developing models that support our current definition of *distant publishing*. These models are illustrated in several use cases, outlined below.

- **Extracted Features Worksets** - HTRC expects this concept to be further refined as we move toward the second round of HTRC Advanced Collaborative Support grants which will be funded in summer 2016. Our most progressive case for distant publishing at this point is leveraged through the publication and release of our main extracted features workset. The current workset is a prototype based on the 4.8 million volume public domain collection from the HTDL. Through 2016-17 this workset will be redefined to include more of the HTDL collection. From this initial workset publication we have seen further refinements of the workset by scholars such as Ted Underwood (Underwood et al. 2013), Colin Allen (Murdock, Zeng, and Allen 2016), and Matthew Wilkens (Wilkens 2013).
- **HT+Bookworm** - The HathiTrust+Bookworm (HT+BW) project (2016) presents textual content through interactive visualization. Whereas HT+BW has previously been used in standalone contexts with pre-determined metadata, currently HT+BW is enabling scholars to analyze custom personal collections from within the larger corpus and the use of HT+BW as a supplement to other uses of the HTRC. This concept could eventually become a new possibility for derived workset publication in its own right.
- **HTRC Workset Ontology** - Currently in development, the HTRC Workset Ontology is part of a collections data model by the Workset Creation for Scholarly Analysis project (HTRC 2016), an HTRC research initiative funded by the Andrew W. Mellon Foundation. The resulting HTRC Workset data model is designed to aid humanities scholars by helping them to describe selected portions of the HTDL corpus that serve as the objects of their research. The resulting worksets are persistent, citable, and can be assessed by other scholars for reuse in additional research processes.

Conclusion

Today's digital scholars are embracing new opportunities to explore their disciplines through the type of enhanced computational analysis that the HTRC provides. As the Center works to define emerging possibilities within the context of non-consumptive research, distant publishing will enable us to engage with the community of open data and open software publishers to ensure that our collections are accessible, open and available for the next generation of distant readers and their plans for new forms of scholarship.

Acknowledgment

The author would like to thank the Executive Management Team of the HTRC, J. Stephen Downie Co-Director, Beth A. Plale Co-Director, Beth Naymachchivaya, and John M. Unsworth, and all of the staff of the HathiTrust Research Center and the HathiTrust Digital Library for their contributions to the tools and services that make the concepts in this paper and the research of our users possible.

License

Bibliography

Courant, P. N. (2009). “The Stakes in the Google Book Search Settlement”. *The Economists Voice* 6 (9). Walter de Gruyter GmbH. doi:10.2202/1553-3832.1665.

Zeng, J., Ruan, G., Crowell, A., Prakash, A., and Plale, B. (2014). “Cloud Computing Data Capsules for Non-Consumptiveuse of Texts”. In *Proceedings of the 5th ACM Workshop on Scientific Cloud Computing - ScienceCloud 14.* Association for Computing Machinery (ACM). doi:10.1145/2608029.2608031.

Jett, J., Cole, T. W., Maden, C., and Downie J. S.. (2016). “The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections”. *Journal of Open Humanities Data* 2 (March). Ubiquity Press Ltd. doi:10.5334/johd.3.

HathiTrust Digital Library. (2016). “HathiTrust Research Center Non-Consumptive Use Research Policy.” https://www.hathitrust.org/htrc_ncup.

Downie, J. S., Capitanu, B., Underwood, T., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C. (2015). “Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain Volumes”. HathiTrust Research Center. doi:10.13012/j8td9v7m.

Downie, J. S., Underwood, T., Capitanu, B., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C. (2015). “Word Frequencies in English-Language Literature 1700-1922 (0.2)”.

HathiTrust Research Center. doi:10.13012/J8JW8BSJ.

Moretti, F. (2000). “Conjectures on World Literature”. *New Left Review* 1 (January): 57–58. https://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature.

Moretti, F. (2013). *Distant Reading*. Verso. http://www.worldcat.org/oclc/813931586.

Underwood, T, Black, M. L., Auvil, L., and Capitanu, B. (2013). “Mapping Mutable Genres in Structurally Complex Volumes”. In *2013 IEEE International Conference on Big Data*. Institute of Electrical
& Electronics Engineers (IEEE). doi:10.1109/bigdata.2013.6691676.

Murdock, J., Zeng, J., and Allen, C. (2016). “Towards Cultural-Scale Models of Full Text”. *Arxiv.org*. Arxiv.org. http://arxiv.org/abs/1512.05004.

Wilkens, M. (2013). “Literary Geography at Corpus Scale”. In *Proceedings of Digital Humanities 2013*. Alliance of Digital Humanities Organizations. http://dh2013.unl.edu/abstracts/ab-139.html.

Organisciak, P., Bhattacharyya, S., Auvil, L., Unnikrishnan, L., Schmidt, B., Shamim, M., McDonald, R., Downie, J., Aiden, E. (2016). “Adding Flexibility to Large-Scale Text Visualization with HathiTrust+Bookworm”. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 854-856. http://dh2016.adho.org/abstracts/179.

HTRC. (2016). “Workset Creation for Scholarly Analysis – A HathiTrust Research Center Project Funded by the Andrew W. Mellon Foundation.” http://worksets.htrc.illinois.edu/worksets.

Online Shadow Libraries and the Future of Humanities Scholarship

Stephen Reid McLaughlin
stephen.mclaughlin@utexas.edu
University of Texas Austin, United States of America

In a publishing landscape long challenged by small audiences, tightening library budgets, and steady growth in the number of monographs produced each year, potentially illegal text sharing has become an everyday practice for many humanities scholars and students, both those with access to conventional distributions channels (libraries, online databases, inter-library loan) as well as independent scholars and those at institutions with limited library resources. Whereas many researchers in medicine and natural science disciplines have embraced open access publication practices, both in open-licensed journals and via centralized preprint repositories, access to scholarship in humanities fields poses greater challenges.

Due to humanities monographs’ limited commercial value, many titles are only commercially available for a short window. As Borgman has noted, “[l]iterature in the humanities goes out of print long before it goes out of date” (2007: 241), while older works from university presses are “virtually entombed” (McGann, 2014: 133) in the stacks of academic libraries. Meanwhile, factors including rising journal prices and an increase in the number of monographs published each year have led libraries to purchase proportionally fewer monographs than in past decades (Thompson, 2005: 103–6). In turn, humanities scholars’ reputation for “unreadable complexity” (Eve, 2014: 22) may be exacerbated by the unavailability of published research. By contrast, informal circulation can provide a double benefit to scholars, allowing them to secure the prestige of publishing with a name-brand university press while reaching a larger audience than would be possible in print alone (Hall 2013). In this paper, I draw on public metadata sets, download logs from Sci-Hub (Bohannon and Elbakyan, 2016), and interviews with operators and users of illicit text sharing sites to examine their place in contemporary humanities scholarship.

A decade ago, illicit text collections, also known as shadow libraries (Liang 2012; Bodó 2015, 2016), were limited in size and existed on the margins of academic culture. Today, one can find millions of books and tens of millions of journal articles spread among Library Genesis, Aaaaarg, and Sci-Hub, as well as in niche collections such as Memory of the World and Monoskop Log. And yet, these sites’ coverage of literary works and literary scholarship is significantly shallower than is the case for published research in medicine, engineering, and the social sciences.

Most shadow libraries are messy, ad hoc affairs, composed of digital objects in a range of formats and quality levels, drawn from a pre-existing ecosystem of interpersonal text sharing among scholars and students. Metadata for the 1.5 million documents in Library Genesis, which is freely available, exhibits what we might call bounded messiness. A tabular dataset riddled with missing values, text encoding quirks, and duplicate entries, it clearly would not pass muster in an academic library setting. However, the simplicity and flexibility of this system make it well-suited for a text collection compiled, maintained, and mirrored by a culturally diverse community of participants. In this case, a just-good-enough database supports the goals of inclusiveness and replication by others.

In interviews with scholars who use shadow libraries, I have observed a wide range of positions with respect to copyright law and the business of scholarly publishing. While some hope for a future in which copyright is abolished and publishers are driven out of business, many are essentially satisfied with the scholarly publishing ecosystem as it stands today. Most use library resources and purchase physical books, turning to shadow libraries to fill the gaps and evaluate texts’ relevance and quality. Nearly all respondents said they prefer to read printed texts when the option is available. Moreover, if the cost of these sites’ continued existence is that they remain culturally marginal, my participants see this as an acceptable tradeoff.

Several interview participants described their personal document management schemes, an age-old form of scholarly labor that typically remains invisible to outsiders. What is new, however, is that some avid collectors are systematically compiling clean copies of their personal libraries, along with extensive metadata and searchable full-text indexes. They then share these collections among colleagues, either on hard drives or in private online repositories. One respondent described this text curation practice as both a pedagogical strategy and an attempt to shape the scholarly canon in his field of study.

There are no easy answers to the problems shadow libraries pose for publishers and university libraries, leading Bodó (2016) to suggest we should "bet on all horses," throwing support behind academic publishers and shadow libraries alike. For now, exploring how where shadow libraries come from, how they are designed, curated, and maintained, and what their futures may hold — including the eventual, inevitable, closure of individual sites — may help us understand current academic culture and possible future models for humanities research and scholarly culture more broadly.

Bibliography

Bodó, B. (2016). "The Knowledge Singularity." RADiCAL.PiRATiCAL symposium, Malmö, Sweden, May 2016. https://radical.piratical.cryptonomic.net/archives/theknowledgesingularity.

Bodó, B. (2015). "Libraries in the Post-Scarcity Era." In **Porsdam, H.** (ed.), *Copyrighting Creativity: Creative Values, Cultural Heritage Institutions and Systems of Intellectual Property*. Farnham, UK: Ashgate, pp. 75–92.

Bohannon, J. and Elbakyan, A. (2016). "Data from: Who's Downloading Pirated Papers? Everyone." http://datadryad.org/resource/doi:10.5061/dryad.q447c.

Borgman, C. L. (2007). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Eve, M. P. (2014) *Open Access and the Humanities: Contexts, Controversies and the Future*. Cambridge, UK: Cambridge University Press.

Hall, G. (2013). "The Unbound Book: Academic Publishing in the Age of the Infinite Archive." *Journal of Visual Culture*, 12(3): 490–507.

Liang, L. (2012). "Shadow Libraries." *e-flux journal* 37.

McGann, J. (2014). *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, MA: Harvard University Press.

Thompson, J. B. (2005). *Books in the Digital Age: The Transformation of Academic and Higher Education Publishing in Britain and the United States*. Cambridge, UK: Polity Press.

Semantic Domains in Picasso's Poetry

Luis Meneses
ldmm@cs.tamu.edu
Texas A&M University, United States of America

Enrique Mallen
mallen.shsu@gmail.com
Sam Houston State University, United States of America

The question of why Pablo Picasso dedicated a considerable amount of his time to writing around 1935 is open to speculation. Many have cited the Spanish artist's emotional crisis, the political turmoil in Europe in the period between the two wars, and the menace of a fratricidal confrontation in Spain as possible causes. All of these views are predicated on an assumed irreducible conflict between visual composition and verbal expression. However, it should be considered that Picasso's interest in alternative methods of expression might have stemmed from his fascination with linguistic structure as a whole during his cubist period. To quote Marie-Laure Bernadac: "I am in complete agreement with [the] linguistics of cubism as a structural language. Picasso is very conscious of the ambivalence of language, the ambivalence of words ... Picasso always played with the ambivalence of words ... The *papiers collés* [may be thought of] as 'proverbs'; that is, as things that take the place of the verb 'to paint' ... The battle between word and image, between art and reality, between different systems of signs ... but one must never forget that Picasso's cubism is two things at once; it's painting and it's language ... His whole life he was obsessed with the relationship between painting and writing ... the battle between word and image, between art and reality" (Rubin et al., 1992). Many other authors have pointed out the parallel development of structural linguistics as delineated by Ferdinand de Saussure and the explorations of the different phases of cubism by Picasso, Braque and Gris.

The connection between his writings and his plastic works has been explored by Cowling (2002), who refers to both his poems and his artworks as executed in a "spider's web" style: "A glance at the manuscripts reveals that he was attentive to the look and lay-out of the pages, relishing the dramatic impact of such things as variations in the size and style of the script, changes in the flow of ink or the thickness of the nib (and the color when he used crayons), contrasts between letters, numbers, dividing lines and the special punctuation marks he favored, different systems for crossing-out and large blots of ink. The calligraphy varies a good deal and is sometimes ornate and the effect handsome and arresting". For a similar position, we may turn to Baldassari

(Picasso and Baldassari, 2005), who poses a clear link between Picasso's visual works and his poetry: "In the decade between 1925 and 1935 ... Picasso continued to pursue cursive linearity in canvases where poetic shapes and ideograms represented bathers and acrobats. The scrolling lines, bursting constellations, curving grids, and broad strokes that cross the surface of his pictorial work then found a new dimension in his poetic writing. Breton fully sensed its importance ... 'This poetry is unfailingly visual in the same way that the painting is poetic'". Earlier, Daix (Daix and Emmet, 1993) had pointed out that "Picasso did not believe in spontaneous poetry – or painting. His attitude was that of a professional: someone who had put written fragments into his paintings and could certainly paint poems as well. The graphism of his letters and the way they were placed on a page were also a deliberate, visual creation". Despite the close correlation between Picasso's poems and his artworks, one cannot deny that Picasso's poetry is essentially verbal, and not conditioned by plastic principles. This is precisely what makes Picasso's poetry so interesting: it provides a window into Picasso's mind that is separate from his own artistic creations.

Our research has taken us through different approaches in order to analyze Picasso's artistic legacy (Meneses et al., 2008a, Meneses et al., 2011) and his poetry. In our study of Picasso's writings, we first created a concordance of all of his poems and plays, separating the poems by language (Spanish and French) (Meneses et al., 2008b). We then designed a dictionary based on this bilingual concordance of poems that was dynamically generated, identifying the stanzas and lines of each poem. A third step led us to provide an English translation to every French and Spanish term, which allowed us to reduce the number of concepts that Picasso works with. This reduction improved clarity for two reasons: First, we eliminated the language distinction (Spanish vs. French); and second, because we merged morphological variation into single units, so that tense/mood alterations for verbs, and gender/number differences for the noun, for instance, were bypassed. As a result, we ended up with three lexical lists: one for Spanish terms, one for French terms, and one with English translations that links the first two lists. The third lexical list may be considered a semantic dictionary specifying concrete concepts in Picasso's writings.

However, we felt that some aspects of Picasso's poetry needed further exploration. In this paper we propose to explore the possibility that Picasso's transition into poetry is simply one more manifestation of his pursuit of alternative approaches to language as a means of representation. In this sense, one thing that remained to be determined was how concrete concepts in both languages cluster into representative semantic categories; and how these categories interact with each other in semantic networks. For this purpose, we have expanded upon our previous efforts by using statistical models and algorithms. More specifically, we have used Latent Dirchlet Allocation: a hierarchical probabilistic generative model that can be used to represent a collection of documents by topics (Blei et al., 2003) to analyze Picasso's poetry. Our analysis demonstrates that topic modeling can highlight patterns and trends in Picasso's poetry that escape other forms of traditional analysis.

We will elaborate on the details of our current analysis using three points. First, it is safe to assume that all poets limit themselves to a number of representational topics as they compose their poems. We know that the European conflict, the crisis in Picasso's own personal relations and the immediate objects in his surroundings constituted the backdrop of many of his compositions. In this respect, topic modeling has allowed us to more precisely delimit the lexical manifestation of these themes, enabling us to see interrelations between words within particular topics.

Secondly, topic modeling has allowed us to see correlations between concrete themes as identified by certain lexical terms and different languages. These correlations are particularly interesting in the case of a bilingual poet like Picasso. For example, psychological concepts–madness, for example– are often handled in French; while more physical references to his immediate surroundings are circumscribed to Spanish terms. However, in cases where both languages are used to communicate a similar concept, Picasso chooses Spanish when he intends to apply a more folkloric tone. Using the English lexical list of terms, we were able to highlight and identify the interconnections between his poems in different languages.

Finally, we have determined that, in Picasso's poems, certain semantic domains are predominant in each of the two languages he used – Spanish and French. For instance, Picasso is more inclined to refer to food items and everyday objects in his Spanish poems, which thus provides a clear reflection of his physical environment and of the harsh economic situation of this time. On the other hand, his French poems concentrate on more abstract concepts involving politics, religion and sexuality, which may be attributed to the influence of French Surrealist writers.

To summarize, in this paper we propose to analyze the ways in which concepts in Picasso's poems and plays cluster into semantic categories; and in turn, how these categories interact with other concepts within a complex semantic network. Furthermore, through the use of statistical models we have been able to identify and pinpoint representative themes and correlations across different languages. Although topic modeling can point us to patterns and interconnections within Picasso's writings, accurately characterizing the nature of these relationships remains a challenge.

Bibliography

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research,* 3, 993-1022

Daix, P. & Emmet, O. (1993). *Picasso: Life and art*, Thames and Hudson.

Cowling, Elizabeth (2002). *Picasso: style and meaning*. London/New York, Phaidon.

Meneses, L., Furuta, R. & Mallen, E. (2008a). Exploring the Biography and Artworks of Picasso with Interactive Calendars and Timelines. *Digital Humanities 2008*, Oulu, Finland. 160 - 162

Meneses, L., Monroy, C., Furuta, R. & Mallen, E. (2011). Computational Approaches to a Catalogue Raisonné of Pablo Picasso's Works. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis.*

Meneses, L., Monroy, C., Mallen, E. & Furuta, R. (2008b). Picasso's Poetry: The Case of a Bilingual Concordance. *Digital Humanities 2008.* Oulu, Finland.

Picasso, P. & Baldassari, A. (2005). *The Surrealist Picasso*, Flammarion-Pere Castor.

Rubin, W., Varnedoe, K., Reff, T., Cottington, D., Fry, E. F., Poggi, C., Krauss, R. & Bois, Y. A. (1992). *Picasso and Braque: A Symposium*, Museum of Modern Art.

Shelf life: Identifying the Abandonment of Online Digital Humanities Projects

Luis Meneses
ldmm@uvic.ca
Texas A&M University, United States of America

Richard Furuta
furuta@cse.tamu.edu
Texas A&M University, United States of America

The Internet, along with the advent of online technologies has provided researchers with greater opportunities to collaborate and create a myriad of digital projects. Taking our research group as an example, we have collaborated in creating online art catalogs (Meneses et al., 2011), interfaces for visualizing and creating poetry (Meneses et al., 2013), and tools for analyzing and exploring Shakespeare's plays (Meneses et al., 2015, Meneses et al., 2016a). However, the convenience and familiarity of computer networks makes us forget (or overlook) that there is a certain fragility associated with our online tools. In turn, this fragility threatens the completeness and the sustainability of our work over time.

Nowadays, a large portion of the research carried out in the digital humanities includes an online project as one of its components. In turn, these digital projects can be catalogued as distributed resources, which implies that the administrative control of information related to a topic may be spread across online resources and/or collections maintained by multiple scholars in different institutions. This administrative decentralization can lead to changes in content that are often unexpected by a researcher.

These unexpected changes can be caused by different factors or circumstances. Changes can occur because of deliberate actions on part of the collector – for example, reorganization of the structure of the collection, switching to a different content management system, or changing jobs and institutions. Changes might also be due to unexpected events – earthquakes, power outages, disk failures, – or may be due to other uncontrollable factors – death, seizure of computers by law enforcement, or termination of the services from an Internet Service Provider (Mccown et al., 2009).

Over time, great strides have been made to harness and manage the fragility of online resources. Klein and Nelson argued that digital documents do not disappear from the Web, but leave artifacts that can be used to reconstruct them (Klein et al., 2011). Bar–Yosseff et al. carried out experiments to measure the decay of the Web (Bar-Yossef et al., 2004). SalahEldeen determined that nearly 11% of shared resources will be lost one year after being published and that this decay will continue at a 0.02% rate per day (Salaheldeen and Nelson, 2012). Nevertheless, and despite these previous efforts, managing and characterizing change in online environments is a complex problem.

Recently, our research group has been focusing on analyzing the perceptions of change in distributed collections (Meneses et al., 2016b). However, we believe that the inherent characteristics of online digital humanities projects present an interesting (and unique) area for inquiry for two reasons. First, the research aspect of digital humanities projects hinders our previous approaches – as our methods for identifying change in the Web do not fully apply. And second, digital humanities projects have a limited useful life – which is accompanied by research from primary investigator, which may or may not be indicated by updates in the project's content and tools. We have seen many cases of successful projects in digital humanities (that fulfill their original objectives and achieve their expected level functionality) that interestingly become abandoned at some point in time. Examples of abandoned successful projects include the Cervantes Project and the TAMU Herbaria Project. This abandonment might be caused by a different set of reasons –which are not often apparent to its users– such as loss of funding, change in personnel or simply decay in interest. We believe that all these reasons are worthy of study.

All this reasoning led us to formulate the following question: When can online digital humanities projects be considered abandoned? In this paper, we propose to present a study on the persistence and average lifespan of online projects in the digital humanities. More specifically, we will elaborate on their reliance on distributed resources and methods for measuring their shelf life: the average length of time that a digital project can endure without updates until it can ultimately be considered abandoned by its researcher.

Furthermore, we believe that "abandonment" is not necessarily a sufficient designation —as there are different nuances involved. We will proceed to elaborate on them using one of our online projects as an example. Digital

Acting Parts is an online project that encourages active reading and memorization, which in turn leads to a better understanding of Shakespeare's plays. The project has been active since 2013, but online development has shifted to a set of different processes that are carried out behind the scenes. Consequently, the project's online presence has not been updated for some time now (we estimate that it has been at least a year). However, the online tools are quite stable at this point. In this specific case, the lack of updates and new content is not a signal of abandonment. This is clear example of why the rules for traditional websites do not fully apply and new metrics are needed to identify issues concerning online projects in the digital humanities.

Our study is an attempt to categorize change in a very specific domain. As an attempt of categorization, determining the degree of abandonment affecting a digital project over time is a difficult task. A Web resource may gradually degrade from being correct to one that is still of some use by providing access to related information or information about the institution to contact for more information. Abandonment can also be hinted by changes in Web servers, directory structures, etc., which may cause Web requests to still result in a successful responses from a server, yet provide no valid information to the requestor. Based on our findings, we approximate the average shelf life to 5 years, which aligns with reports from previous work (Goh and Ng, 2007).

Additionally, our study will touch upon on potential strategies for the archival and the long-term preservation of abandoned digital online projects. It is important to highlight that different levels of preservation and curation are needed among digital projects. Historically, preservation efforts have been primarily concerned with maintaining the primary artifacts in collections; relegating descriptive metadata to a lesser level of importance. There is an underlying notion that descriptive metadata is static: requiring minimal resources to maintain and consequently making it easier to preserve. However, our previous work (Meneses et al., 2016b) has shown us that this is not the always the case.

To summarize, in this paper we propose to identify indicators of the abandonment of digital humanities projects – as well as identifying their average lifespan. Digital online projects in the humanities have unique characteristics that make them impervious to the metrics that used in the Web as a whole. In our opinion, these unique characteristics make them worthy of study. In the end, the purpose of our study is to gain a better understanding of digital humanities projects, their lifespan and formulate better strategies for their long-term preservation.

Bibliography

Bar-Yossef, Z., Broder, A. Z., Kumar, R. & Tomkins, A. (2004). Sic transit gloria telae: towards an understanding of the web's decay. *Proceedings of the 13th international conference on World Wide Web,* 2004 New York, NY, USA. ACM.

Goh, D. H. L. & Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, 58, 15-24.

Klein, M., Ware, J. & Nelson, M. L. (2011). Rediscovering missing web pages using link neighborhood lexical signatures. *Proceedings of the 11th annual international ACM/IEEE Joint Conference on Digital libraries,* 2011 Ottawa, Ontario, Canada. ACM.

Mccown, F., Marshall, C. C. & Nelson, M. L. (2009). Why web sites are lost (and how they're sometimes found). *Communications of the ACM*, 52, 141-145.

Meneses, L., Estill, L. & Furuta, R. (2015). Digital Acting Parts: Learning and Understanding Shakespeare's Plays*. Joint CSDH/SCHN & ACH Digital Humanities Conference 2015.* Ottawa, Canada.

Meneses, L., Estill, L. & Furuta, R. (2016a). This was my speech, and I will speak it again": Topic Modeling in Shakespeare's Plays*. Joint CSDH/SCHN & ACH Digital Humanities Conference* 2016. Calgary, Canada.

Meneses, L., Furuta, R. & Mandell, L. (2013). Ambiances: A Framework to Write and Visualize Poetry. *Digital Humanities 2013.* University of Nebraska–Lincoln.

Meneses, L., Jayarathna, S., Furuta, R. & Shipman, F. (2016b). Analyzing the Perceptions of Change in a Distributed Collection of Web Documents. *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, Halifax, Nova Scotia, Canada. ACM, 273-278.

Meneses, L., Monroy, C., Furuta, R. & Mallen, E. (2011). Computational Approaches to a Catalogue Raisonné of Pablo Picasso's Works. *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis.*

Salaheldeen, H. M. & Nelson, M. L. (2012). Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? *Proceedings of Theory and Practice of Digital Libraries 2012*, 2012 Paphos, Cyprus.

Decolonizing Knowledge Structures in Open Access and Scholarly Publishing

Nirmala Menon
nmenon@iiti.ac.in
Indian Institute of Technology, India

One of the agenda questions for the Internet Researchers Conference (IRC) 2017 scheduled to be held at IIIT, Bengaluru India is: "What research tools and infrastructures are needed to study, document, annotate, analyze, archive, cite, and work with (in general) digital objects, especially those in Indic languages?" However, before we get to this core question in the research arena we need to confront the fundamental gap in the conversation– that there are as yet very few digital objects that are created, developed and preserved in Indian languages. Shodhganga, a digital repository for dissertations and theses from universities in India, is a national digitization project that has now grown substantially as a scholarly resource. Similarly, we have the National Digital Library, a project that plans to make curricular access electronic. However, both extraordinary projects are still limited to works that are primarily written in English. This is especially unfortunate in literature research. Literature departments across universities in India have a robust scholarship agenda and have some niche publications that focus on the works of literatures of the different languages as recognized in the 8th schedule of the constitution of India. What is missing is a bridge or a platform that enables a conversation between the scholarships in these languages. Certain university department structures– those lacking a department of Literature hospitable to comparatist conversations– have to take some responsibility for this lack of conversation. There are complicated economic and social factors beyond the scope of this discussion that are at play in this disproportionate representation of Indian literatures. It suffices to say, we need specific digital tools that recognize and fill this gap.

At the Open Access Scholarly Publishing project at IIT Indore, we have tried to identify and fulfil what we see as two scholarly imperatives: 1) a commitment to open access environment as crucial for research accessibility and 2) providing a platform for multilingual scholarly production with an emphasis on Indian languages. The Digital Humanities and Publishing Research Group at IIT Indore has a two-pronged approach to what we see as a fundamental infrastructure problem: the absence of a polyphonic platform for a scholarly kitchen! The publishing project KSHIP was born from deliberation over these crucial issues challenging scholarship publishing in India.

KSHIP or Knowledge Sharing in Publishing recognizes this gap and advocates a philosophy of open access as crucial to a cooperative global knowledge production network. In its initial pilot/launch phase, KSHIP includes 1) a project of database development of scholarship and criticism in Indian languages. (we are starting with three languages, with the eventual goal of making KSHIP a crowd sourced platform with more languages to be added soon), and 2) we will be a consciously multilingual publishing house soliciting (and specifically targeting) scholarly monographs and translations in Indian languages, while also inviting scholars to host journals in multiple languages. In this session we would like to discuss some of the technological challenges of the project and possible solutions with an emphasis on collaborations across institutions and scholars in India.

The database development project addresses the first issue of dissemination and access to scholarship in regional language literatures in India. We have started a pilot with four languages with citation data from three different languages being uploaded. This will truly be a crowd sourced project in the future and we are in discussion with journals/publishers about uploading data to the platform. At least in the beginning, this platform will also be a part of KSHIP. We are looking for scholars from different languages to contribute to and be a part of this project. The project will be collaborative in the largest sense of the term.

The project also includes a prototype of a translation plug-in tool. This is a crowd-sourcing translation tool developed as part of an UG project and directed by myself and my colleague from Computer Science, Dr Abhishek Srivastava. We hope to finesse the initial prototype of the tool over the coming months (and may even have a final prototype ready by the conference in Montreal!). The translation tool is designed to slice a scholarly paper/article so that one paper can potentially be translated by more than one person, and then moderated and collated at the publisher’s end. There are several challenges posed by such a crowd sourced model. These include, among others, issues of author/translator attribution, for which we are currently discussing and thrashing out different options. The interface will also offer the option of solicited translation, especially in the case of literature scholarship. We will experiment with the crowd-sourced translation using scientific and engineering essays.

In this short paper I will discuss: 1) The infrastructure gap in scholarly publishing in India, 2) the need for this to be addressed with a focus on multilingual publishing, and 3) the technological challenges specific to an ambitious long-term project. These questions will inevitably lead to discussions on digital preservation and the transforming and transformative role of technology in that process.

ELLE the EndLess LEarner: Exploring second language acquisition through an endless runner–style video game

Donald F. Merritt II
don.merritt@ucf.edu
University of Central Florida, United States of America

Emily Kuzneski Johnson
emily.johnson@ucf.edu
University of Central Florida, United States of America

Amy Larner Giroux
amy.giroux@ucf.edu
University of Central Florida, United States of America

Games have been used to help people learn throughout history (Vankúš, 2005). Video games are interactive and encourage active learning (Domínguez, Saenz-de-Navarrete, de-Marcos, Fernández-Sanz, Pagés, & Martínez-Herráiz, 2013; Watson, Mong, & Harris, 2011) more so than learning via lecture or textbook reading. The interactive nature of games also makes them engaging to the player. Student engagement is an essential component of learning (Fredricks, Blumenfeld, & Paris 2004; Hall, Ramsay, & Raven, 2004; Kahu, 2013). Most scholars agree that games, when designed well, can increase student learning and retention (Ricci, Salas, & Cannon-Bowers, 2009) as well as motivation (Gee, 2003; Fullerton, 2014; Eichenbaum, Bavelier, & Green, 2014).

Sykes and Reinhardt (2013) define game-based second language learning and teaching as the "use of games and game-inclusive synthetic immersive environments" (p. 5) that are designed specifically for second language learning contexts. They also state that few game-based spaces are available specifically for second language learning purposes. Scholars have noted the unintentional effect on second language learning that commercial games can have, for example, children in countries where English is not commonly spoken have been observed as acquiring knowledge of the language as a means to play the game (Sørensen & Meyer, 2007). Other research has focused on second language-learning interactions that learners form in multi-player video environments, including those associated with video game chat rooms (Ryu, 2013). Although some video games that target language learning have emerged, such as *Croquelandia, Language Island, Mentira, MIDDWorld Online,* and *Zon* (Sykes & Reinhardt, 2013), much research remains to be done in the efficacy of this game genre. The central question guiding our research is: How can a videogame best be designed to effectively enhance student second language acquisition?

To answer this question, we turn to Ellis's (1985) recommendations to facilitate language learning: the quantity of input directed at the learner, the learner's perceived need to communicate in the second language, and the here-and-now principles. Current second language pedagogy also stresses that learning should be authentic and goal-oriented. Drawing on socio-cultural theory, researchers have linked second language acquisition with a "language game" (Lantolf, 1997), explaining that play can allow learners to rehearse linguistic forms they already know and to experiment with new input in a low-pressure environment. The game we are developing will satisfy Ellis's requirements for the quantity of input, the necessity for competency in the language (to advance in the game), and the "here-and-now" principle, in addition to authenticity and goal-orientation.

This session will explain and demonstrate *ELLE the EndLess LEarner*, a game prototype designed to enable the study of secondary-language acquisition through an endless runner game platform. The system is designed so that different game features, specifically auditory, visual, and textual cues, can be modified easily by a researcher and the efficacy of each studied in relation to language acquisition and the game platform. The platform has been designed in this way to allow for research into the impact of these different cues and input methods on the activities being studied.

The style of *ELLE* is an "endless runner," which means that the player's avatar is continuously in motion, "running" through the game environment. For example, the iOS game *Temple Run 2*, continuously moves the player forward, without allowing the avatar to stop. The only controls available to the player are turning right or left, and jumping over or ducking under obstacles. The player cannot control the speed of motion though it can be manipulated through in-game objects. This game style results in a fast pace requiring players to react rapidly, and it has great potential to increase student motivation and engagement in the types of common sense practice exercises that help solidify vocabulary acquisition. *ELLE* is being designed in conjunction with modern language researchers and is grounded in the scholarship of language learning theories and evidence-based pedagogical methods.

We chose an endless runner style videogame to address a number of practical concerns, as scholars assert that the connection between game mechanics and intended argument must be intentional (Bogost, 2007). First, this popular style of game is intuitive and easy to play. Next, we believe the endless runner style, being a casual game, will be less intimidating for learners who do not identify as "gamers." The fast pace of this style requires rapid recall and provides immediate feedback in a low-consequence environment that lends itself to repeated play and therefore repeated reinforcement of the content. Endless runner games limit player choices within the game, which allow for an increased emphasis on the language terms being practiced. Finally, this prototype, initially focusing on vocabulary acquisition, lays the groundwork for future iterations of the game as an on-demand language practice tool that will have the ability for instructor modification of content.

We have chosen Portuguese as the initial language for *ELLE* to teach, as it is one of the most important world languages today. As the sixth most spoken language in the world, the Portuguese language is a gateway to the cultural and economic opportunities of Brazil and the rest of the Portuguese-speaking world. From Portugal to the emerging economies of Africa, the Portuguese language has over 200 million native speakers, is an official language in nine countries, and is spoken on three continents. The US government considers Portuguese to be a language of "critical need," especially because of the nation's trade relations with Brazil.

ELLE is being designed as a game system that can continue to be studied and built upon over time. The team's initial goal is for *ELLE* to function as the center of a variety of studies to increase understanding of second language acquisition. *ELLE* is being designed using the *Unreal Engine* and readily available tools and tutorials, to reduce costs and to allow for increased focus on the variety of cue types that are of interest to the research team. This game prototype style, with its manipulable components, will allow the team to quantify the effects of different game features and player actions. Additionally, the components can be used to increase accessibility and to adapt the game based on player needs. This empirical data can then be leveraged and further explored to increase the efficacy of language learning games.

Our research will work to identify what motivates and engages SLA learners, relying on self-determination theory (Ryan, Rigby, & Przybylski, 2006) to ensure an appropriate balance of intrinsic and extrinsic motivational strategies. The design team's collaboration with modern language instructors has identified a lack of postsecondary level SLA supplemental tools that students find motivating; therefore, this prototype has the potential to inform improvements to future tools for second language instruction. This game is intended to act as a supplement to instruction and other classroom activities, rather than a replacement. Because student motivation is such a major factor in learning, the learning research undertaken as part of this project will also examine student engagement using the model of mastery and performance goal orientation (De Clercq, Galand, and Frenay, 2014). Mastery of second language concepts will motivate some students while others may focus on their performance against their peers. In addition to testing the game feature of textual, visual, and auditory cues, the team is also investigating the impact of game mechanics such as leader boards on student motivation.

The team's long-term goal is to create a robust database-driven game that can be easily customized to teach vocabulary words / phrases / pictograms from any semiotic domain, regardless of language or subject, particularly at the primary and secondary school levels. *ELLE* will be leveraged to research socio-cognitive variables of domain language acquisition. Such socio-cognitive variables include questions about how the gaming environment influences learners' motivation and how the interaction within the multimodal space of a video game could be affected by different aspects of identity, for example, gender. Potential semiotic domains could even include STEM subjects such as chemistry and biology. In other words, the proposed project will allow the investigative team not only to develop an effective game through which to introduce a second or additional language but, ultimately, to allow the team to explore both cognitive and social aspects of language acquisition.

Bibliography

Bogost, I. (2007). *Persuasive games: The expressive power of videogames.* Cambridge, MA: MIT Press.

De Clercq, M., Galand, B., & Frenay, M. (2014) Learning processes in higher education: Providing new insights into the effects of motivation and cognition on specific and global measures of achievement. In D. Gijbels, V. Donche, J.T.E. Richardson, and J.D. Vermunt (Eds). *Learning Patterns in Higher Education: Dimensions and research perspectives.* (pp. 140-162). London: Routledge.

Domínguez, A., Saenz-de-Navarrete, J., de-Marcos*, L., Fernández-Sanz, L., Pagés, C., Martínez-Herráiz, J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education, 63*(2013), 380–392.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford: Oxford University Press.

Eichenbaum, A., Bavelier, D., & Green, C. S. (2014). Video games: Play that can do serious good. *American Journal of Play, 7*(1), 50-72.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*, 59-109.

Fullerton, T. (2014). What games do well: Mastering concepts in play. In W. G. Tierney, Z. B. Corwin, T. Fullerton, & G. Ragusa (Eds.) *Postsecondary play: The role of games and social media in higher education* (pp. 125-145). Baltimore: Johns Hopkins University Press.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Hall, M., Ramsay, A., & Raven, J. (2004). Changing the learning environment to promote deep learning approaches in first-year accounting students. *Accounting Education, 13*(4), 489-505.

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in higher education*, *38*(5), 758-773.

Lantolf, J. P. (1997). The function of language play in the acquisition of L2 Spanish. In A. Pérez-Leroux & W. R. Glass (Eds.), *Contemporary perspectives on the acquisition of Spanish* (pp. 3-24). Somerville, MA: Cascadilla Press.

Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (2009). Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology, 8*(4), 295-307.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344-360.

Ryu, D. (2013). Play to learn, learn to play: Language Learning through gaming culture. *reCALL 25*(2), 286-301.

Sykes, J. & Reinhardt, J. (2013). *Language at play: Digital games in second and foreign language teaching and learning*. New Jersey: Pearson Education.

Sørensen, B. H., & Meyer, B. (2007). Serious games in language learning and teaching–a theoretical perspective. In *Proceedings of the 3rd international conference of the digital games research association* (pp. 559-566).

Vankúš, P. (2005). History and present of didactical games as a method of mathematics' teaching. *Acta Didactica Universitatis Comenianae-Mathematics*, *5*, 53-68.

Watson, W. R., Mong, C. J., & Harris, C. A. (2011). A case study of the in-class use of a video game for teaching high school history. *Computers & Education, 56*(2), 466–474.

An Automated Approach to Model the Transformation Process of the Reuse in Bernard de Clairvaux: How Do Lexical Resources Help?

Maria Moritz
mmoritz@etrap.eu
University of Goettingen, Germany

Marco Büchler
mbuechler@etrap.eu
University of Goettingen, Germany

Abstract

To fortify the research of automated, historical text reuse detection, it is necessary to investigate the way in which a text is reused (e.g., verbatim, paraphrased) in order to understand the broader context of a reuse. Our long-term goal is to build a formal theory behind reuse transformations. We have previously investigated two datasets of Bible reuse to analyze how reuse is modified and how linguistic resources support this. In this work, we investigate the ratio of non-literal text reuse, and we measure to which extent the Ancient Greek WordNet—which also contains Latin WordNet—and BabelNet can support identifying lexical relations in Latin reuse excerpts. In doing so, we also show the lack and need of resources for ancient data.

Introduction

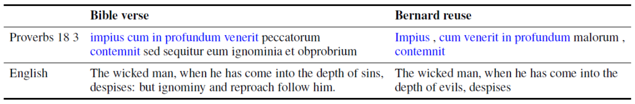
The automated detection of historical text reuse is still in its early stages. To reinforce its research, it is necessary to investigate the way a text is reused in order to understand the broader context. Here is where the necessity of lexical resources supporting this task comes in, especially when a text is non-literally reused, and words are substituted with semantic equivalents, such as synonyms or other semantically similar words. Our long-term goal is to formally model reuse transformations. The analysis of the amount and type of substitutions of words with lexically related words enables insights into how text is reused. Applying these insights into future development of detection methods helps to improve them. We have previously investigated two datasets of Bible reuse, trying to understand how reuse is modified (when operations are performed on word pairs) and how linguistic resources support this task. To achieve this, we need to study more and different cases of reuse. In this short paper, we propose and report on work that extends the number of reuse excerpts we investigated in previous work (Moritz et al., 2016), and take another linguistic resource, BabelNet into account. We aim at investigating the current state of lexical resources' support for a Latin reuse dataset. We compare the support we can get from an additional lexical resource to previous results. Specifically, we investigate BabelNet (BN) (Navigli and Ponzetto, 2012), a multiple resource network pulling from different sources, and we compare the reuse detection support (how many words are covered) between BN and Ancient Greek WordNet (AGWN) (Bizzoni et al., 2014), which also contains Latin WordNet (Minozzi, 2009). Both are recently developed resources and the most common for the Latin language. BabelNet is produced from a range of different, contemporary sources, such as Wikipedia and Wikidata. We are interested in the extent to which BabelNet is able to cover words and relations from an ancient reuse dataset. We are especially curious about what words are still supported by current resources. Our ultimate goal is to simulate a transformation process that also supports non-literal reuse. This can help to model the changes that were applied to an ancient text during its reuse history.

Background

The field of automatically detecting historical text reuse is still in its early stages. To date, Büchler (2013) combines state-of-the-art NLP techniques to address reuse detection scenarios for historical texts, ranging from near copies to text excerpts with a minimal overlap, using a method, which selects n-grams from an upfront pre-segmented corpus. While the approach can discover historical and modern text reuse language-independently, it requires a minimal text similarity. Recognizing modified reuse is difficult in general. Alzahrani et al. (2012) study plagiarism detection techniques, such as n-gram-, syntax-, and semantics-based approaches. However, as soon as reused text is modified (e.g., word substitution), most systems fail. Finally, lexical resources support the identification of relationships between words, but they are not free from issues (Jing, 1998) that can appear when they are used to adapt a general lexicon to a specific domain (Miller et al., 1990).

Data

Our dataset contains excerpts from twelve works—mainly sermons and treaties (Literature)—and two work collections—sermons and letters—from the Latin writer Bernard of Clairvaux (c.f., Moritz et al., 2016). All those texts come from the *Sources Chrétiennes* collection (c.f., Mellerin, 2014) (changes in format and orthography may be inserted by the editor). The Biblindex project (Mellerin, 2014) extracted over thousand Bible reuse exerpts from these works, each of which points to a Bible verse. We use the Latin Bible from Biblindex, called *Biblia sacra juxta vulgatam versionem* (Weber R., 1969) to link the excerpts to the respective Bible verses. We come up with 1,128 unique reuse-to-bible-verse pairs. Table 1 shows one example.



	Bible verse	Bernard reuse
Proverbs 18 3	impius cum in profundum venerit peccatorum contemnit sed sequitur eum ignominia et obprobrium	Impius , cum venerit in profundum malorum , contemnit
English	The wicked man, when he has come into the depth of sins, despises: but ignominy and reproach follow him.	The wicked man, when he has come into the depth of evils, despises

Table 1: Example of reuse

Methodology

We use AGWN, which is automatically constructed from Greek-English digitized lexicons, which again were provided by the Perseus Project (Crane, 1985) and also aligns to Minozzis Latin WordNet (Minozzi, 2009). BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic network that integrates lexicographic and encyclopedic knowledge from WordNet (Fellbaum, 1998), Wikipedia, and others. We further use lemma lists from the Biblindex project, as well as the Latin lemma list from the Classical Language Toolkit (CLTK), which is available in the online GitHub repository of the CLTK (Johnson et al., 2014 2016), to increase the hit rate when querying both resources.

To model the transformation in-between two text excerpts, we define replacement operations (OPs) (see Table 2) that represent the transformation of a reuse to the Bible verse it refers to, as well as an algorithm that identifies those operations between word pairs of a reuse and a Bible verse in a prioritized order. Our algorithm first finds all possible operations for a reuse word, and then applies the most literal operation using the counterpart Bible verse word, which fulfills this operation. This means that if no perfectly or lemmatized matching word is found, relationships of semantic closeness (such as synonyms) for a given word are retrieved. We call the group of semantic operations non-literal operations (c.f., Table 3). We apply our algorithm (which identifies the operations) on Bernard's reuse, first using the relationships queried from AGWN and second, using BabelNet. Afterwards, we show which operations are identified, and calculate a support value for both processes.

operation	description	example
NOP(reuse_word, orig_word)	Original and reuse word are equal.	*NOP(maledictus,maledictus)*
upper(reuse_word, orig_word)	Word is lowercase in reuse and uppercase in original.	*upper(filio,Filio)*
lower(reuse_word, orig_word)	Word is uppercase in reuse and lowercase in original.	*lower(Gloriam, gloriam)*
lem(reuse_word, orig_word)	Lemmatization leads to equality of reuse and original.	*lem(penetrat, penetrabit)*
repl_syn(reuse_word, orig_word)	Reuse word replaced with a synonym to match original word.	*repl_syn(magnificavit, glorificavit)*
repl_hyper(reuse_word, orig_word)	Word in bible verse is a hyperonym of the reused word.	*hyper(cupit, habens)*
repl_hypo(reuse_word, orig_word)	Word in bible verse is a hyponym of the reused word.	*hypo(dederit, tollet)*
repl_co-hypo(reuse_word, orig_word)	Reused word and original have the same hyperonym.	*repl_co-hypo(magnificavit, fecit)*
lemma_missing(reuse_word, orig_word)	Lemma unknown for reuse or original word.	*lemma_missing(tentari, inlectus)*
no_rel_found(reuse_word, orig_word)	Relation for reuse or original word not found in word net.	*no_rel_found(gloria, arguitur)*

Table 2: List of operations and corresponding examples (cf. Moritz et al., 2016)

Results

Table 3 shows the identified operations. Using AGWN, we encounter a high ratio of synonyms (repl_syn), a lot of co-hyponyms and a significant number of hyperonyms and hyponyms. With BabelNet these figures are about a tenth as high. Table 3 shows that the values for **NOP**, **lower** and **lem** (matching words, and words with same lemma) slightly differ in-between both word nets. This is caused by a design decision of our algorithm, which pragmatically permits to reassign a word when it already was used in an association with an earlier word.

	literal				non-literal				unclassified		
	NOP	upper	lower	lem	repl_syn	repl_hyper	repl_hypo	repl_co-hypo	no_rel_found	lem_missing	total
AGWN	4521	1	396	770	397	125	124	316	2470	450	9570
BN	4524	1	397	771	25	21	36	112	3233	450	9570

Table 3: Absolute numbers of operations identified

Fig. 1 shows that AGWN outperforms BabelNet in identifying semantic relations, which represent non-literal text reuse, because these ratios are much lower for BabelNet than for AGWN. We further encounter three significant descents: between 0% and 10%, 30% and 40%, and 50% and 60%. Looking into samples deeply, we find three patterns: i) the more semantic related words are replaced in a reuse, the more likely it is an allusion or analogy, and the less paraphrased or verbatim it is; ii) short allusions are better covered by the Latin synsets than paraphrases with a high ratio of semantic related words; iii) paraphrases with a high literal ratio are covered best. We summarize that both word nets cover paraphrased reuse to a certain extent of replaced words, and AGWN better identifies allusions.

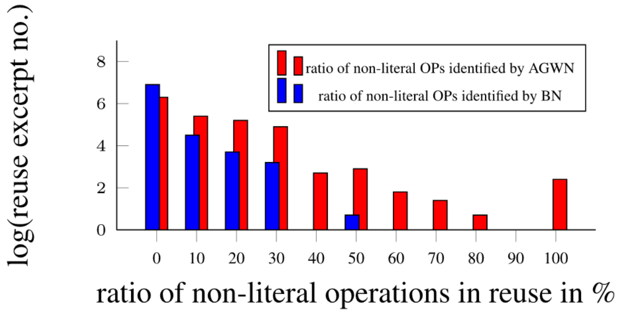


Figure 1: Ratio of non-literal (semantic) operations, aggregated in 10%-steps in relation to the whole reuse length. The reuse number is displayed logarithmically due to clarity reasons.

Lastly, we calculate a support value, which determines the ratio of non-literal operations (c.f., Table 3) compared to them including unsuccessful resource look-ups (no_rel_found) in both, AGWN and BN. For AGWN this value is about 28%, for BabelNet about 6%. Both values are to be understood as lower bounds, because often there is no reasonable relationship in-between two words.

Even if BN coverage is poor, its results tell us, which words of a dataset of medieval, Biblical Latin and Latin of the church fathers are prevailed in a current resource. Some examples are words such as **gloria** (glory) (contained in 17 synsets), **corona** (crown) (contained in 10 synsets), or **nemo** (nobody) (contained in 4 synsets).

Conclusion

We identified the ratio of non-literal reuse in a Latin dataset and showed the support of two lexical resources. Our results show that language resources for Latin reuse are limited and that only a small part of the required coverage is supported. This result raises awareness for the lack of resources for ancient data, despite the growth of language resources for modern languages. Our future work includes refining our operation set, analyzing more languages, increasing the size of our datasets, and investigating probability measures for those data in lexical hierarchies. Since lexical resources will never completely cover the vocabulary at hand, we further consider the application of a form of word embedding.

Acknowledgements

We thank Laurence Mellerin for providing the dataset we used, and for advice on its content. Our work is funded by the German Federal Ministry of Education and Research (grant 01UG1509).

Bibliography

Alzahrani, S. M., Salim, N., and Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. Trans. Sys. Man Cyber Part C, 42(2):133–149.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., and Crane, G. (2014). The making of Ancient Greek WordNet. In Proceedings of the Ninth International Conference on

Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. European Language Resources Association (ELRA).

Büchler, M. (2013). Informationstechnische Aspekte des Historical Text Re-use (English: Computational Aspects of Historical Text Re-use. Ph.D. thesis, Leipzig University, Germany).

Crane, G. (1985). Perseus digital library. http://www.perseus.tufts.edu/hopper/.

Fellbaum, C. (1998). WordNet: An electronic lexical database: Bradford book.

Jing, H. (1998). Usage of WordNet in natural language generation. In Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (COLING-ACL'98). Columbia University Academic Commons.

Johnson, K.P., Burns P.J., Hollis, L., Pozzi, M., Shilo, A., Margheim, S., Badger, G., and Bell, E. (2014–2016). Cltk: The classical language toolkit. https://github.com/cltk/cltk.

Mellerin, L. (2014). New ways of searching with Biblindex, the online index of biblical quotations in early Christian literature. In Claire Clivaz, Andrew Gregory, and David Hamidovic, editors, Digital Humanities in Biblical, Early Jewish and Early Christian Studies, chapter 11, pages 175–192. Brill, Leiden.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K.J. (1990). Introduction to WordNet: An on-line lexical database. International Journal of Lexicography (special issue), 3(4):235–312.

Minozzi, S. (2009). Innsbrucker Beitrge zur Sprachwissenschaft, volume 137, chapter The Latin WordNet Project, pages 707–716. Institut für Sprachen und Literaturen der Universität Innsbruck, Innsbruck.

Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y., and Büchler, M. (2016). Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In Empirical Methods in Natural Language Processing (EMNLP'16), Austin, TX, USA. Association for Computational Linguistics.

Navigli, R., and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. 193:217–250.

Weber, R., Gribomont, J., Fischer, B., Eds. (1969) 1969, 1994, 2007. Biblia Sacra Juxta Vulgata Versionen. Deutsche Bibelgesellschaft.

Novel Approaches to Research and Discover Urban History

Sander Münster
sander.muenster@tu-dresden.de
Technische Universität Dresden, Germany

Kristina Friedrichs
kristina.friedrichs@tu-dresden.de
Technische Universität Dresden, Germany

Cindy Kröber
cindy.kroeber@tu-dresden.de
Universität Würzburg, Germany

Jonas Bruschke
jonas.bruschke@tu-dresden.de
Universität Würzburg, Germany

Frank Henze
frank.henze@btu-cottbus.de
Technische Universität Dresden, Germany

Florian Niebling
florian.niebling@uni-wuerzburg.de
Universität Würzburg, Germany

Abstract

The research group on four-dimensional research and communication of urban history (HistStadt4D) aims to investigate and develop methods and technologies to transfer extensive repositories of historical media and their contextual information into a three-dimensional spatial model, with an additional temporal component. This will make content accessible to different target groups, researchers and the public, via a 4D browser. A location-dependent augmented-reality representation can be used as an information base, research tool, and to communicate historical knowledge. The data resources available for this research include extensive holdings of historical photographs of Dresden, which have documented the city over the decades, and digitized map collections on the Deutsche Fotothek (German photographic collection) platform. These will lay the foundation for a prototype model which will give users a virtual experience of historic parts of Dresden.

Introduction

Imagine you're exploring the historic center of a city with its impressive town houses, churches and monuments. What if you could just use your mobile to find out about the historic buildings around you, with detailed visual information about how they were built and the story behind them, mak-ing history come alive before your eyes?

Photographs and plans are an essential source for historical research (Burke, 2003, Paul, 2006, Wohlfeil, 1986, Pérez-Gómez and Pelletier, 1997) and key ob-jects in eHumanities (Kwastek, 2014). Numerous digital image archives, containing vast numbers of photographs, have been set up in the context of digitization projects. These extensive repositories of image media are still difficult to search. It is not easy to identify sources relevant for research, analyze and contextualize them, or compare them with the historical original. The eHumanities research group HistStadt4D, funded by the German Fed-eral Ministry of Education and Research (BMBF) until July 2020, is investigating and developing methods and technologies for this. The junior research group consists of 14 people – including 4 post-doctoral and 4 PhD researchers. Since a focal interest is to comprehensively investigate how to enhance accessibility of large scale image repositories, researchers and research approaches originate from the humanities, geo- and information technologies as well as from educational and information studies. In contrast to adjacent projects dealing primarily with large scale linked text data as the Venice Time Machine project (2017), sources addressed by the junior group are primarily historical photographs and plans. Historical media and their con-textual information are being transferred into a 4D – 3D spatial and temporal scaled - model to support research and education on urban history. Content will be made accessible in two ways; via a 4D browser and a location-dependent augmented-reality representation. The prototype database consists of about 200,000 digitized historical photographs and plans of Dresden from the Deutsche Fotothek (German photographic collection).

Key Aspects

Usage scenarios and research values

Digital image repositories meet a wide range of needs, from research in humanities and information technologies, through museum contexts and library studies to tourist applications (Münster, 2011). Architectural historians have developed various methods of analyzing both preserved and never-built or destroyed structures in chronology and context (Brassat and Kohle, 2003). Style analysis, iconographic approaches, and art sociological methods all address structural historical questions. The technological possibilities of digital image repositories allow architecture historians to draw on a much larger stock of material, and to process and evaluate this from new perspectives. In addition, innovative software tools can be used to locate sources temporally and spatially, or to support dating, stylistic criticism, authorizations or archaeological investigations (Verstegen, 2007). Depending on the user group, a number of contradictory requirements must be met. Historical researchers, for example, need to be able to compare and contextualize sources (Münster et al., 2015, Brandt, 2012, Wohlfeil, 1986), and to trace the relationship between source and representation (Favro, 2006, Niccolucci and Hermon, 2006). This includes identifying formal patterns, singularities, and discontinuities in architecture and cityscape. This raises a host of questions: How do buildings and cities change over time? In which contexts, such as political or formal developments, does a historical cityscape evolve? What similarities can be found between objects in terms of construction standards and requirements, building codes, regional, temporal or personal tastes and styles?

The research group will address these and many more questions in a specific project on the interdependence between urban development and urban photography.

Creating targeted tools for working with image repositories

An adjacent task will be to perform a systematic survey of the needs of users of image repositories, whose findings will be used to conceptualize technological support options. As historic images, objects and information are increasingly being digitized on a massive scale, more content becomes available for investigation; more cross-analyses are possible; more knowledge is collected, structured and shared (Schuller, 2009). The new scale of research and information retrieval creates many new challenges. Many scholars note that online searching for images and information is "counter-productive" due to the amount of irrelevant data retrieved or their limited technical abilities (Beaudoin and Brady, 2011). Access and efficient data retrieval is inhibited for a variety of reasons. The degree of search expertise is as important as the functionalities and usability of the platform (Kemman et al., 2014). A lot of the existing tools of research programs and applications stem from computer science and do not necessarily meet the needs of humanities scholars (Dudek et al., 2015). Users need efficient search and filter functions, an intuitive software interface and navigation system (Barreau et al., 2014). Appropriate documentation through metadata plays an important role in ensuring sustainability (Bentkowska-Kafel et al., 2012, Maina and Suleman, 2015). In contrast, users expect an intuitive and feasible introduction to the topic and data (Maina and Suleman, 2015) with options to find out more as required. The simplest way to link and contextualize visual information is to use highlighted keywords as hyperlinks in texts and captions. Data interaction and processing tools are also essential for research (Webb and O'Carroll, 2013, Hecht et al., 2015).

Photogrammetric methods of visual knowledge generation

A possible technological basis for creating access to large scale image repositories is the spatial and temporal aggregation of data, in this case historical photographs in a 4D model. The potential of photographic images ranges from pure documentation in archaeology and monument preservation, through image interpretation, for example for damage documentation, to the production of complex 3D models for archaeological investigations (Bührer et al., 2001). Geometrical reconstruction from historical photographs is based on photogrammetry. Information from multiple 2D images is used to acquire 2D and 3D object geometries and have frequently been applied on historical and measurement images (cf. Wiedemann et al., 2000, Bräuer-Burchardt and Voss, 2001, Henze et al., 2009, Siedler et al., 2011). Since some decades, traditional analytical photogrammetry has increasingly been complemented by digital image processing and analysis. The elaborate process of manual image analysis can be largely automated, resulting in large image collections from which geometric information can be generated automatically (Pomaska, 2011). To date, automated photogrammetric methods are generally used primarily to evaluate current, mostly digital images. So far, this has rarely been done for historical images, as it involves specific challenges. Scanned analogue records usually have unknown camera metrics, missing or minimal object information and low radiometric and geometric resolution. Our aim is to develop application-oriented tools for photogrammetric analysis of historical photographs, to integrate them into the process of historical image analysis (Fig 1), and to create a spatial relationship to today's situation.

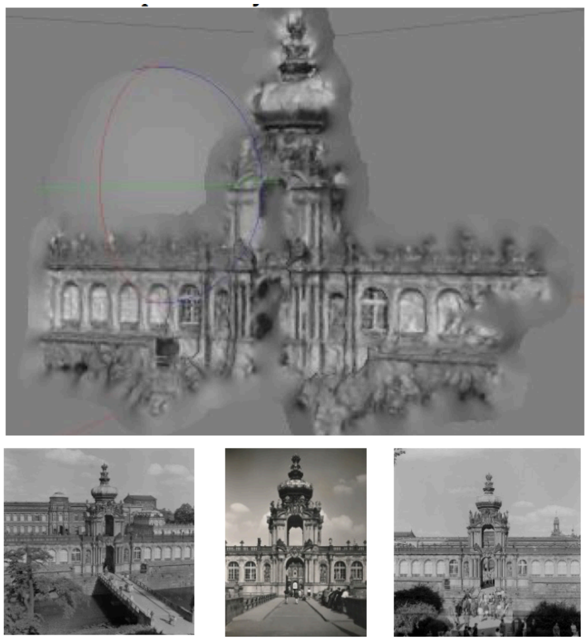


Figure 1. 3D model based on current photographs and historical photographs (proof-of-concept)

Augmented reality

The prototype 4D model, and the 4D historical photographs, drawings, plans, and information within it, will be made accessible via a location or context related information access as augmented reality (Münster and Niebling, 2016). This technology has gained importance in the last few years and undergone extensive testing (Livingston et al., 2008, Zöllner et al., 2010, Walczak et al., 2011, Chang et al., 2015, Chung et al., 2015). Augmented reality describes the enrichment of the real world through virtual data, which can include 3D models, texts, pictures, films or audio data.

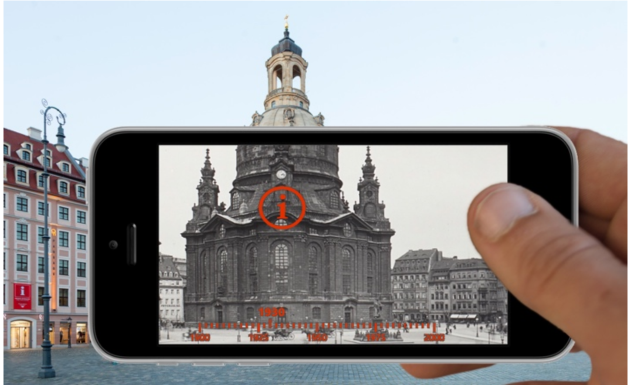
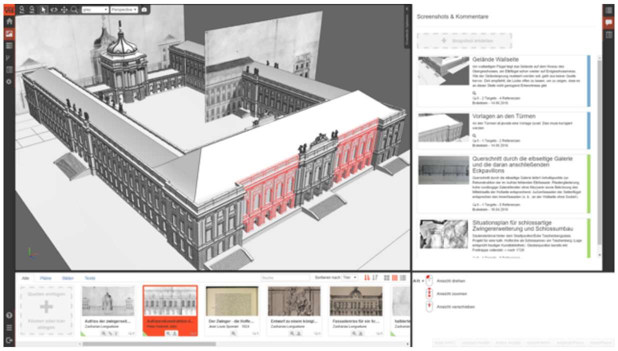


Figure 2. Augmented-reality representation of a cityscape (mockup)

Enriching the reality or replacing parts of reality can help to bridge the cognitive gap between our daily life experience of a cityscape and its depiction in historical photographs (Niebling et al., 2008). In the historical context, the viewer is able to interactively capture visual and textual information about objects in their historical spatial reference system (Ridel et al., 2014). Our investigation will focus on the accessibility of historical data: How can interaction with virtual buildings be designed? Which metaphors can be used? How can augmented reality support educational and research settings?

4D browser



Screenshots & Kommentare
Snapshot erstellen
Gelände Wallseite
Vorlagen an den Türmen
Querschnitt durch die elbseitige Galerie und die daran anschließenden Eckpavillons
Situationsplan für schlossartige Zwingererweiterung und Schlossumbau
Alt +
Ansicht drehen
Ansicht zoomen
Ansicht verschieben

Figure 3. 4D-Browser (prototype)

As an alternative path, the 4D model will also be accessible via a 4D browser interface for spatially and temporally located searches in media repositories. An basing application prototype of a research platform for 3D reconstruction projects is in development was developed during a master thesis (Bruschke, 2015), employing approaches, such as semantic data linking and visualization of temporally and spatially arranged information (Gouveia et al., 2015). Since the prototype has focused on individual building complexes, the 4D browser application has to visualize an entire city model, which also changes constantly over time. Moreover, a visual interface is proposed to make additional information accessible, such as the current and original location of the depicted object. Further features intended to support scholarly users of the prospected platform are image rectification tools and overlays combining several pictures from different periods which can shed light on changes in a building. Statistical analyses of photographed objects over time may provide information on a building's significance. Last but not least, the application should be intuitive to operate for a heterogeneous user group (Warwick, 2012).

Summary

As a result of huge and concerted digitization efforts, extensive digital repositories of historical photographs have been created in the past few decades. This volume of data presents a major challenge to support search, access and information enrichment for users. In August 2016, the HistStadt4D research group started examining scientific methodological requirements and intuitive user interfaces for dealing with massive media repositories from a multi-disciplinary perspective.

Bibliography

Deutsche Fotothek (n.d.) [Online]. Available: http://www.deutschefotothek.de/ [Accessed 9.5.2014].

Deutsches Dokumentationszentrum für Kunstgeschichte - Bildarchiv Foto Marburg (n.d.) [Online]. Available: http://www.fotomarburg.de/ [Accessed 9.5.2014].

École Polytecnique Fédérale de Lausanne. (2017. The Venice Time Machine [Online]. Available: http://vtm.epfl.ch/page-109337.html, 11.01.2017 [Accessed].

Barreay, J-B., Gaugne, R., Bernard, Y., Le Cloirec, G., and Gouranton, V. (2014). Virtual reality tools for the West Digital Conservatory of Archaeological Heritage. Proceedings of the 2014 Virtual Reality International Conference.

Beudoin, J. E., and Brady, E. (2011.) Finding Visual Information: A Study of Image Resources Used by Archaeologists, Architects, Art Historians, and Artists. Art Documentation, 30, 24-36.

Bentowska-Kafel, A., Denard, H., and Baker, D. (2012). Paradata and Transparency in Virtual Heritage, Burlington, Ashgate.

Brandt, A. V. (2012). Werkzeug des Historikers, Stuttgart [u. a.], Kohlhammer.

Brassat, W., and Kohle, H. (2003). Methoden-Reader Kunstgeschichte. Texte zur Methodik und Geschichte der Kunstwissenschaft, Köln.

Bräuer-Burchardt, C., and Voss, K. (2001). Facade Reconstruction of Destroyed Buildings Using Historical Photographs. In: Albertz, J. (ed.) Proceedings of the XVIII. International CIPA Symposium 2001, IAPRS, Vol. XXXIV, Part 5/C7, 2001.

Bruschke, J. (2015). DokuVis – Ein Dokumentationssystem für Digitale Rekonstruktionen (Master thesis). Master thesis, HTW Dresden.

Bührer, T., Grün, A., Zhang, L., Fraser, C. & Rüther, H. (2001). Photogrammetric Reconstruction and 3D Visualization of Bet Gorgis, a Rock-hewn Church in Ethiopia. In: Albertz, J. (ed.) Proceedings of the XVIII. International CIPA Symposium 2001, IAPRS, Vol. XXXIV, Part 5/C7, 2001.

Burke, P. (2003). Augenzeugenschaft. Bilder als historische Quellen, Berlin.

Chang, Y.-L., Hou, H.-T., Pan, C.-Y., Sung, Y.-T. and Chang, K.-E. (2015). Apply an Augmented Reality in a Mobile Guidance to Increase Sense of Place for Heritage Places. Educational Technology & Society, 18, 166-178.

Chung, N., Han, H. and Joun, Y. (2015). Tourists' intention to visit a destination: The role of augmented reality (AR) application for a heritage site. Computers in Human Behavior, 50, 588-599.

Dudek, I., Blaise, J.-Y., De Luca, L., Bergerot, L. & Renaudid, N. (2015). How was this done? An attempt at formalising and memorising a digital asset's making-of. Digital Heritage, 2, 343-346.

Favro, D., (2006). In the eyes of the beholder. Virtual Reality re-creations and academia. In: Haselberger, L., Humphrey, J. & Abernathy, D. (eds.) Imaging ancient Rome: Documentation, visualization, imagination: Proceedings of the 3rd Williams Symposium on Classical Architecture, Rome, 20.- 23. 5. 2004. Portsmouth: Journal of Roman Archaeology.

Gouveia, J., Branco, F., Rodrigues, A. & Correia, N. (2015). Travelling Through Space and Time in Lisbon's Religious Buildings. In: Guidi, G., Scopigno, R., TorresJ. C. & Graf, H. (eds.) 2nd International Congress on Digital Heritage 2015. Granada.

Hecht, R., Meinel, G. & Buchroithner, M. F. (2015). Automatic identification of building types based on topographic databases - A comparison of different data sources. International Journal of Cartography, 1, 18-31

Henze, F., Lehmann, H. & Bruschke, B. (2009). Nutzung historischer Pläne und Bilder für die Stadtforschungen in Baalbek / Libanon. Photogrammetrie - Fernerkundung - Geoinformation, 3, 221–234.

Kemman, M., Kleppe, M. & Scaglolia, S. (2014). Just Google It. Digital Research Practices of Humanities Scholars. Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities. Sheffield: HRI Online.

Kwastek, K. (2014). Vom Bild zum Bild. Digital Humanities jenseits des Texts (Keynote). 1. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2014). Passau.

Livingston, M. A., Bimber, O. & Saito, H. (2008). Proceedings of the 7th IEEE International Symposium on Mixed and Augmented Reality. Cambridge, UK, Piscataway, IEEE Xplore.

Maina, J. K. & Suleman, H. (2015). Enhancing Digital Heritage Archives Using Gamified Annotations. In: Allen, B. R., Hunter, J. & Zeng, L. M. (eds.) Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015. Proceedings. Cham: Springer International Publishing.

Münster, S. (2011). Entstehungs- und Verwendungskontexte von 3D-CAD-Modellen in den Geschichtswissenschaften. In: Meissner, K. & Engelien, M. (eds.) Virtual Enterprises, Communities & Social Networks. Dresden: TUDpress.

Münster, S.., Jahn, P.-H. & Wacker, M. (2015). Von Plan- und Bildquellen zum virtuellen Gebäudemodell. Zur Bedeutung der Bildlichkeit für die digitale 3D-Rekonstruktion historischer Architektur. In: Ammon, S. & Hinterwalder, I. (eds.) Bildlichkeit im Zeitalter der Modellierung. Operative Artefakte in Entwurfsprozessen der Architektur und des Ingenieurwesens. München: Wilhelm Fink Verlag.

Münster, S. & Niebling, F. (2016.) HistStadt4D - Multimodale Zugänge zu historischen Bildrepositorien zur Unterstützung stadt- und baugeschichtlicher Forschung und Vermittlung. Digital Humanities im deutschsprachigen Raum (DHd) 2016. Duisburg: nisaba verlag.

Niccolucci, F. & Hermon, S. (2006). A Fuzzy Logic Approach to Reliability in Archaeological Virtual Reconstruction. In: Niccolucci, F. & Hermon S. (eds.) Beyond the Artifact. Digital Interpretation of the Past. Budapest

Niebling, F., Griesser, R. T. & Woessner, U. (2008). Using Augmented Reality and Interactive Simulations to Realize Hybrid Prototypes. Advances in Visual Computing, 4th International Symposium, ISVC 2008 (Proceedings, Part I). Las Vegas, NV.

Paul, G. (2006). Von der Historischen Bildkunde zur Visual History. Visual History. Ein Studienbuch. Göttingen.

Peréz–Gómez, A. & Pelletier, L. (1997). Architectural Representation and the Perspective Hinge, Cambridge, London, University Press.

Pomaska, G. (2011). Zur Dokumentation und 3D-Modellierung von Denkmalen mit digitalen fotografischen Verfahren. In: Heine, K., Rheidt, K., Henze, F. & Riedel, A. (eds.) Von Handaufmaß bis High Tech III - 3D in der historischen Bauforschung. Mainz: Verlag Philipp von Zabern.

Ridel, B., Reuter, P., LaViole, J., Mellado, N., Couture, N. & Granier, X. (2014). The Revealing Flashlight: Interactive Spatial Augmented Reality for Detail Exploration of Cultural Heritage Artifacts. J. Comput. Cult. Herit., 7, 1-18.

Schuller, G. (2009). Designing universal knowledge, Baden, Lars Müller Publishers.

Siedler, G., Sacher, G. & Vetter, S. (2011. Photogrammetrische Auswertung historischer Fotografien am Potsdamer Stadtschloss. In: Heine, K., Rheidt, K., Henze, F. & Riedel, A. (eds.) Von Handaufmaß bis High Tech III - 3D in der historischen Bauforschung. Mainz: Verlag Philipp von Zabern.

Verstegen, U. (2007). Vom Mehrwert digitaler Simulationen dreidimensionaler Bauten und Objekte in der architekturgeschichtlichen Forschung und Lehre. Vortrag am 16.3.2007. XXIX. Deutscher Kunsthistorikertag, 2007 Regensburg.

Walczak, K., Cellary, W. & Prinke, A. (2011). Interactive Presentation of Archaeological Objects Using Virtual and Augmented Reality. In: Jerem, E., Redö, F. & Szevereni, V. (eds.) On the Road to Reconstructing the Past. Proceedings of the 36th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA). Budapest: Archaeolingua.

Warwick, C. (2012). Studying users in digital humanities. In: Warwick, C., Terras, M. & Nyhan, J. (eds.) Digital Humanities in Practice. London: Facet Publishing,.

Webb, S. & O'Carroll, A. (2013). Digital Heritage Tools in Ireland - a Review. Papers of Cultural Heritage, Creative Tools and Archives, 26.–27.06.2013, National Museum of Denmark, Copenhagen,

Wiedemann, A., Hemmleb, M. & Albertz, J. (2000.) Reconstruction of historical buildings based on images from the Meydenbauer archives. International Archives of Photogrammetry and Remote Sensing, XXXIII, 887–893.

Wohlfeil, R. (1986). Das Bild als Geschichtsquelle. Historische Zeitschrift, 243, 91–100.

Zöllner, M., Becker, M. & Keil, J. (2010). Snapshot Augmented Reality - Augmented Photography. In: Artusi, A., Joly-Parvex, M., Lucet, G., Ribes, A. & Pitzalis, D. (eds.) 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST 2010). Paris: Eurographics Association

Digital Analysis of the Literary Reception of J.W. von Goethe's *Die Leiden des jungen Werthers*

Sandra Murr
sandra.murr@ts.uni-stuttgart.de
Universität Stuttgart, Germany

Florian Barth
florian.barth@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Introduction

The impact of Johann Wolfgang von Goethe's epistolary novel *Die Leiden des jungen Werthers* marks a singular moment in German literary history, as it sparked a remarkable critical and productive literary reception. After the publication in 1774, so-called *Wertheriaden* — literary adaptations that are formally, thematically and structurally guided by Goethe's novel — appeared in various languages. These texts form our corpus for the Werther Project, which is part of the interdisciplinary DH Project CRETA of the University of Stuttgart.

Literary studies have mainly concentrated on individual texts of this literary reception and on the individual genres to which the adaptations belong. Our research approaches focus on a macro-perspective to recognize recurring structures, elements, motifs and character constellations in all of the *Wertheriaden* with the help of computer-assisted analysis. From this perspective, we want to determine how closely the literary adaptations are based on Goethe's original text *Werther.*

Corpus

Our Werther corpus contains about 150 German and 30 English texts from different genres and literary periods — mainly (epistolary) novels, dramas and poems. The rationale for our corpus selection is based on *Werther's* publication in 1774 and its translation into French (1776) and English (1779). In the last two decades of the 18th century no other work of German literature was ever been translated or adapted in English as often as Goethe's novel.

To establish comparative parameters for such a heterogeneous corpus (which also allow an analysis across all literary genres), we concentrate on specific features which are present in Goethe's original work. Among these discernible and typical characteristics (Martens 1985) are the triangular relationship of the three main characters, the concept of the "Werther character" (emotional, artistic), the monoperspectival narration, the role of nature (with motifs such as "Herbst", i.e. autumn), the subject's relation to society, the so-called "sickness unto death" ("Krankheit zum Tode") with the protagonist's subsequent suicide, as well as structural, stylistic and linguistic similarities (Horré 1997). These reference points will be applied leading to a large scale comparison.

One approach is to compare the different texts on the basis of network and lexical information. In this context, we are concentrating on the visualization of the typical character network of Goethe's original text and whether this triangular constellation reoccurs in the *Wertheriaden.*

Character Networks

The analysis of character constellations is an essential part of literary studies. It shows the dynamic structure of the interactions of the central characters in the text and characterizes them in relation and contrast to each other (Pfister 1982). In the last few years, there has been a growing interest in graph-theoretic visualization and analysis of social networks in literary texts — especially in dramatic texts (Hettinger et al., 2015, Trilcke 2013, Moretti 2011).

Central character constellations occur in all literary genres and form the basis for conflict constellations, particularly for the static counter-narrative between a protagonist and an antagonist. In Goethe's *Werther*, this typical constellation is extended by another figure — Lotte. She mediates between the emotional protagonist Werther, who is immortally in love with her, and Albert, her fiancée and Werther's adversary with opposite characteristics. A central aspect of the Werther Project is the identification of this triad by visualizing the character constellation in the original text and the literary adaptations, as well as detection of deviations.

Method

For the determination of character relations, name lists were derived from Goethe's epistolary novel and its adaptations in close reading. This lexicon-based approach leads to better results than machine learning techniques for named-entity recognition, since complex entities ("Graf C." or "Frau von S.") which exist in the original text have to be determined without doubt. For each character we paid attention to synonyms (e.g. Lotte, Lottgen, Lottchen) and in case of the Werther character we added the personal pronoun "I", due to the first-person narration. Connections are established when two named entities are located at an adjustable distance of n tokens (normalization, stop word removal or sentence borders can be set as parameters). Based on this simple heuristic approach, the typical triad can be identified in both Goethe's *Werther* (fig. 1) and in Ulrich Plenzdorf's modern adaptation *Die neuen Leiden des jungen W.* (fig. 2) from 1972. Another constitutive element of the plot is the addressee of the messages (Wilhelm/Willi) of the Werther character (Werther/Edgar). With respect to the triangular relationship, we aim to compare the vectors of a single pair, like Werther-Lotte in Goethe's novel, with a corresponding pair, e.g. Edgar-Charlie in Plenzdorf's adaption.

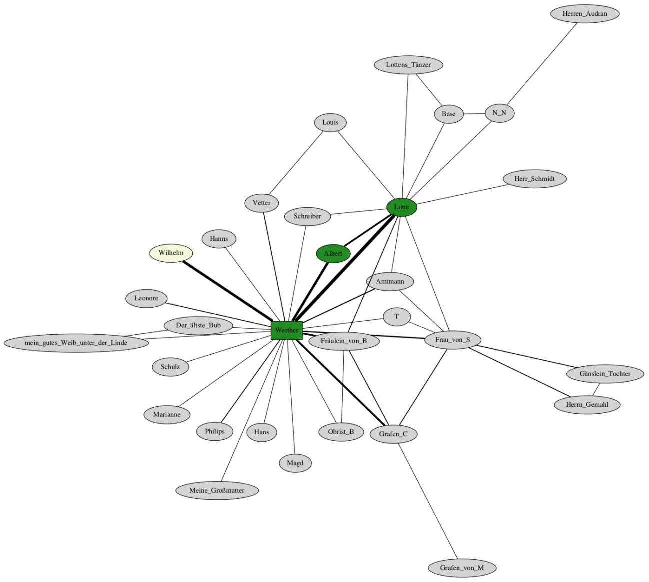


Figure 1: Character network Die Leiden des jungen Werthers (1774); Distance for connections: 8 tokens

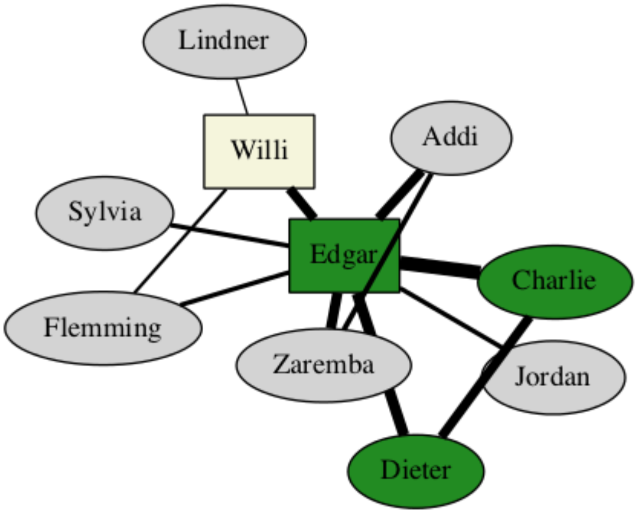


Figure 2: Weighted network of Die neuen Leiden des jungen W. (1972); Distance for connections: 8 tokens

Edgar, the Werther character in Plenzdorf's adaption, takes the central node with the highest degree (the maximum number of connections, cf. fig 3). We noticed a slightly different position of the beloved women in both texts: In Goethe's Werther Lotte appears to be very well connected with other nodes, compared to Charlie in Plenzdorf's novel, who is only in contact with Edgar and his antagonist Dieter (cf. fig. 2).

Die neuen Leiden des jungen W. produces a denser network based on fewer character nodes with more weighted edges. That means Edgar has relatively intense contact to other characters besides his message partner. In contrast, many more characters appear in Goethe's novel, but often only once in connection with Werther, which reduces the network's density.

Average Degree	2.95	2.4
Maximum Degree	21 (Werther)	8 (Edgar)
Density	0.089	0.27

Figure 3: Network measures

Future prospects

The character constellation is determined by an opposition of the sensitive Werther to the rational opponent character Albert, with their love interest Lotte displaying character traits of both. Furthermore, an examination of the relationship pairs with regard to their context is planned.

The usefulness of such an approach is illustrated by the sentence below, which defines the relationship of Lotte and Werther in context:

> **Sie** [Lotte] stand auf ihrem Ellenbogen gestützt und ihr Blick durchdrang die Gegend,

> **sie** sah gen Himmel und auf mich, **ich** [Werther] sah
> ihr Auge tränenvoll, sie legte ihre
> Hand auf die meinige und sagte — **Klopstock!**

The term "Klopstock" is distinctive for the “Werther-Lotte” pair. It reflects their spiritual kinship and characterizes their relationship. However, this expression is found exactly once in Goethe's epistolary novel. Overall, the "Werther-Lotte" pair appears 81 times in the text. Based on these instances we aim to characterize their relationship even with specific terms like "Tanze", "Porträt" or “Klopstock”. This will complete the network visualization of the characters with a description of their individual relations, both in Goethe’s novel and its adaptations.

Acknowledgements

As noted in the Introduction, the Werther Project is part of CRETA, a research center connecting both scholars from Humanities/Social Sciences and Computational Linguistics/Computer Science at the University of Stuttgart. We are grateful to the German Federal Ministry of Education and Research (BMBF) for its generous funding in the years 2016 until 2018 (funding reference number: 01UG1601)

Bibliography

Hettinger, L. et al. (2015): Genre Classification on German Novels, *Proceedings of the 12th International Workshop on Text-based Information Retrieval*, Valencia.

Horré, T. (1997): *Werther-Roman und Werther-Figur in der deutschen Prosa des Wilhelminischen Zeitalters*, St. Ingbert.

Martens, L. (1985): *The Diary Novel,* Cambridge.

Moretti, F. (2011): Network Theory, Plot Analysis, in*: New Left Review* 68: 80-102.

Pfister, M. (1982): *Das Drama. Theorie und Analyse,* München: 232-250.

Trilcke, P. (2013): Social Network Analysis (SNA) als Methode einer textempirischen Literaturwissenschaft, in: Philip Ajouri, Katja Mellmann u. Christoph Rauen (Hgg.): *Empirie in der Literaturwissenschaft,* Münster 2013: 201-247.

A Generic Tool for Visualizing Patterns in Poetry

Onur Musaoğlu
onurmusaolu@gmail.com
Bogazici University, Turkey

Özge Dağ
dagozge@gmail.com
Bogazici University, Turkey

Özge Küçükakça
ozge.kucukakca@boun.edu.tr
Bogazici University, Turkey

Alkım Almıla Akdağ Salah
almilasalah@sehir.edu.tr
Sehir University, Turkey

Albert Ali Salah
salah@boun.edu.tr
Bogazici University, Turkey

Introduction

Visualization of text can be a useful exploration tool for looking at the corpus of a poet, especially when dealing with a prolific author with a large body of output over the years. In this work, we describe a flexible and extensible tool for analyzing the corpus of a poet, and make a case study of Nâzım Hikmet Ran. Since poetry has its own challenges over plain text, we have developed novel ways of visualizing the structure, the rhythm and affective tone of each poem, as well as ways of looking at the continuities (or discontinuities) of features in the entire corpus over the years. The designed system integrates a database for holding meta-information, and a website for creating and linking interactive, parameterized visualizations.

Nâzım Hikmet Ran is one of the most famous poets of Turkey. Although he was a great patriot, he has spent many years in prison and in exile due to his communistic political views. His poetry is translated into more than fifty languages. We believe this tool can be particularly useful to compare different translations of the poet’s work, to see how certain stylistic or semantic features are retained (or lost) during translations.

Related Work

Most poetry visualizations focus on the aesthetics of information rather than the functional aspects. An example is Diana Lange’s visualizations that transform individual poems into beautiful visual displays, resembling flowers. A similar project is Boris Müller's Poetry on the Road, which turns a text into an image through an arbitrary transformation function, for instance by treating every word as a location and creating a heat map of the entire text. The outcomes of these projects do not tell us much about the poets or the œuvres in question.

In contrast to such artful renderings of poems, there are studies that focus solely on the grammatical and structural problems of poetry writing. Such studies rather try to find quantitative ways to analyze poems, enabling a computational approach for the evaluation of technical quality and subtlety of the rhymes (Opara, 2014; Dalvean2015). In the same line of research, there are visualization tools such as Graphwave, SentimentGraph, SentimentWheel and Ambiances (Meneses and Furuta, 2015). Such visualization examples constitute the starting point of our explorations for devising a new visualization system that is both scalable and modular in nature, i.e. a tool that would accommodate different natural language processing (NLP) tools, as well as new visualization techniques.

Methodology

In this section, we briefly describe the database structure, as well as the software tools used to create the system.

Database

After his death, Nâzım Hikmet's collected works appeared in a single volume (Nâzım, 2007). The digital version of this volume is not publicly available, but we received a special permission from the publishers to use this volume.

Nâzım is a poet who paid special attention to the visual structure of his poems and it is imperative to retain this structure as accurately as possible. Consequently, line indentations were kept intact for each line, as well as the fonts of the individual words. We also paid attention to the fact that the collected works included some text written in prose. Thus, the database structure, depicted in Fig. 1, is entirely hierarchical and ordered according to books, works in a book, lines in a work, words in a line, and characters in the words. This may seem to be an extensively elaborate representation, but it allows detailed structural analysis, as well as the analysis of visual and rhythmic features of each work.

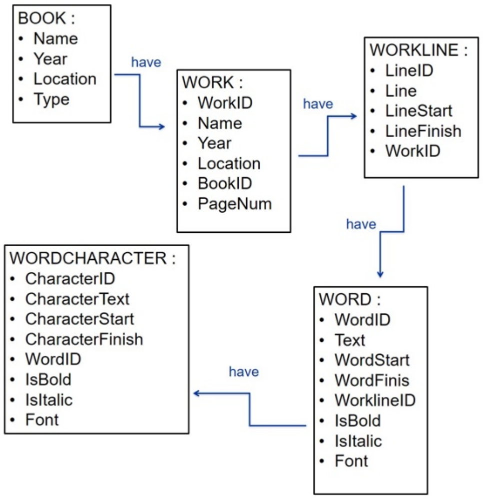


Fig. 1: Database structure.

Software

Since the project involves a dynamic, parametric and interactive system, many software technologies were used. To keep structured data, user data and web page related content, a MySQL database was used. A Turkish-based affect analysis tool was integrated with the system, and Perl was employed to read and parse data for the affect analysis tool. The main programing language of the project is Java and all back end code is also developed in this language. The Spring Framework was used to create the model-view-controller (MVC) structure of the application. In the front end of the application, AngularJS was used to create the MVC structure and to create a single page application. Moreover, to make the application responsive, CSS3 and Bootstrap technologies were used for mobile phone support.

NLP Challenges

In order to parse Nâzım's corpus, a Turkish Morphological Parser and Disambiguator was used (Sak et al., 2008). With the help of this tool, we get part of speech tags of the words, as well as some grammatical information about verbs (i.e. conjugation and tense) and about words' grammatical number. For certain instances, the morphological parser suggests more than one form or number. To solve such problematic instances, an off-the-shelf disambiguator was used. The results of this disambiguation tool suggests the most appropriate form for a given context, which helps in making a decision on the preferred form of a given verb, noun, or pronoun.

The system was enhanced with a text-based affect analysis tool, which returns valence, arousal and dominance values for a given sentence and each word in that sentence (Aydın Oktay et al., 2015). One of the challenges in parsing poems versus prose texts is the lack of a specific notation for indicating the end of a sentence. For the sake of simplicity, each line of a given poem was treated as a sentence, and valence, arousal and dominance values were computed for every word and line individually. These values are stored in the database for fast retrieval.

The System Interface

The generated system works as an interactive visualization tool with a web interface. For the user experience of the web system, a responsive interface is prepared that can even be reached via smart phones. Also, to keep data alive and to allow flexible operations, a single page application is created, with which users can surf between different tabs without losing information. The system design is modular and expandable, as each work unit is separated such that new visualizations can be easily added to the system. The system can also be tailored to visualize a new database easily. The only requirement is that the work of the artist be parsed in the same hierarchical way, and placed on a SQL-capable database.

Visualizations

One of the motivations behind the visualizations is to give information about the analysis on the corpus of Nâzım, and other poets when the database is expanded by the addition of new authors. The tool incorporates a search function, and allows different visualizations to be prepared from the results of the search. Most of the prepared visualizations are interactive charts. They can be used for showing a term's usage over the years, over geographic locations and over publications. The search can be conducted on a collection of works, or in a single work. A separate page was created for searching for a sequence of words, and to prepare comparative visualizations.

We briefly describe two visualizations here to serve as examples. The first visualization is called the "poetry barcode" (see Fig. 2). In this visualization, one poem is visualized, and each line of the poem is represented by a horizontal line. The length and the color of the lines are set according to NLP and affect analysis results, and the lines form six different columns, which show the change of line lengths, usage of active/passive phrases in a line, inflections of words in a line according to person information, as well as valence, arousal, and dominance values of each line.

Nâzım has a lot of stylistic features in his poems. To be able to analyze and extract these features, we have prepared visualizations about the usage of alliteration and his unique verse structure. Alliteration is a stylistic device, in which a number of words, having the same consonant sounds, occur close together in a series. To quantify alliteration, a measure was developed that uses the background frequency of each letter in the poet's corpus. By using a sliding window based evaluation, letter frequencies are calculated, and compared with the base frequency of that letter in the corpus. Fig. 3 illustrates the automatic alliteration detection. Fig. 4 shows a number of additional visualizations in a bird's eye view, including a visualization about passive/active voice usage.

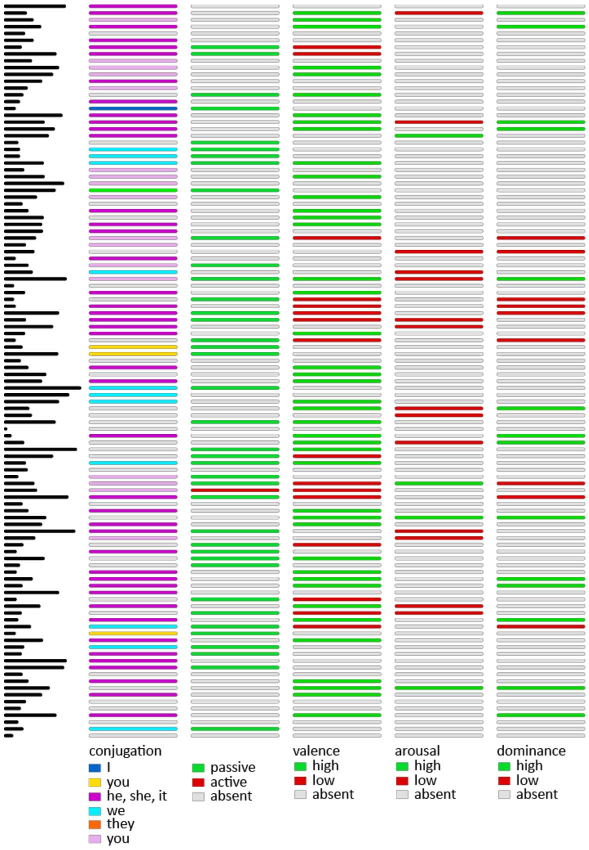
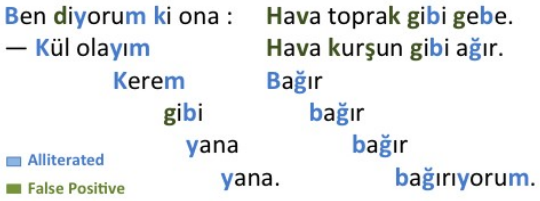


Fig. 2: A poem's barcode, visualizing the structure of the poem together with verb conjugations, passive/active verb usage, and emotional tone.



Ben diyorum ki ona :
— Kül olayım
Kerem
gibi
yana
yana.

Hava toprak gibi gebe.
Hava kurşun gibi ağır.
Bağır
bağır
bağır
bağırıyorum.

Fig. 3: Visualizing alliteration in the poem. Best viewed in color.

Conclusions

A complete web page opened to the wider public is in construction, as it requires some security features due to copyrights of the works in the database. But the system is operational in the offline mode, and already provides many visualization options.

Since the database contains composition years and places for poems (where available), it is possible to search for words that are historically relevant for Nâzım. To give an example, a search for the words "hürriyet," (freedom) and "hapis" (jail) restricted to 1919-1938 and 1938-1963, we can easily see that these words' usages are significantly increasing after 1938, when he was arrested for the first time. Other possible uses include the visualization of words associated with different colors, prominent in his poetry, over the years.

The proposed system keeps data and visualizations separate, but well-connected. This enables the addition of new artists to the proposed system. The tool also can be used as a platform to evaluate poetry translation. We show grammatical and affective features of the words in some visualizations like the "poetry barcode". It is possible to use these visualizations to get an idea about translation quality in literal translations, where the emphasis is on word-for-word translation.

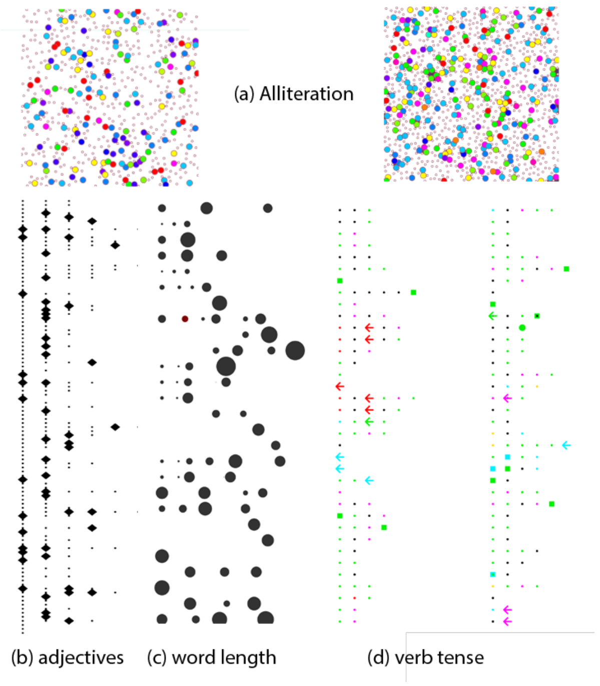


Fig. 4: A bird's eye view of several visualization options in the system. Best viewed in color.

Acknowledgments

We would like to thank Yapı Kredi Yayınları for giving us permission to use the digital version of Nâzım's works. This study is realized in collaboration with Nâzım Hikmet Kültür ve Sanat Araştırma Merkezi.

Bibliography

Aydın Oktay, E., Balcı, K., and Salah, A. A. (2015). Automatic assessment of dimensional affective content in Turkish multi-party chat messages. In *Proc. Int. Workshop on Emotion Representations and Modelling for Companion Technologies*, pages 19–24. ACM.

Dalvean, M. (2015). Ranking contemporary American poems. *Digital Scholarship in the Humanities*, 30(1):6–1

Meneses, L. and Furuta, R. (2015). Visualizing poetry: Tools for critical analysis. *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture*, 3(1).

Nâzım, H. (2007). Bütün şiirleri. *İstanbul: YKY*

Opara, K. R. (2014). Grammatical rhymes in Polish poetry: A quantitative analysis. *Digital Scholarship in the Humanities*, page fqu029.

Sak, H., Güngör, T., and Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in natural language processing*, pages 417–427. Springer.

(Not) Going to the movies: a geospatial analysis of cinema markets in The Netherlands and Flanders (1950–1975)

Julia Noordegraaf
j.j.noordegraaf@uva.nl
University of Amsterdam, Netherlands

Jolanda Visser
j.a.t.visser@uva.nl
University of Amsterdam, Netherlands

Jaap Boter
jaap.boter@vu.nl
VU University, Amsterdam, Netherlands

Daniel Biltereyst
daniel.biltereyst@ughent.be
Ghent University, Belgium

Philippe Meers
philippe.meers@uantwerpen.be
University of Antwerp, Belgium

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, Netherlands

Introduction

Cinema as a new cultural industry at the dawn of the twentieth century has had a significant impact on the social, cultural and economic infrastructures of modernizing societies. Although cinema from its first emergence was widely adopted, the penetration of cinemas on the exhibition market (number of cinemas, attendance figures) has shown remarkable differences between European countries. Considering the density of cinema theatres and attendance figures as a marker for market penetration, in particular The Netherlands and Belgium stand out: while Belgium abounded in cinemas and film attendance, cinema density and attendance in the Netherlands were traditionally low (Biltereyst and Meers, 2007).

Several scholars have attempted to explain these differences in market penetration, yet no comprehensive, satisfactory explanation has hitherto been found. The low cinema density and attendance in The Netherlands has been explained from the organization of society in various religious and ideological 'pillars' (pillarization) that, until at least the 1970s, compartmentalized the social, cultural, economic and political spheres of society but that excluded the cinema market, that operated 'neutrally' (Dibbets, 2006). However, in similarly compartementalized Flanders a pillarized cinema landscape was successfully created (Biltereyst and Meers, 2007). Others have pointed out that class might have been a possible factor (Thissen and Van der Velden, 2009). Finally, the organization and economics of the industry have been identified as influencing the distribution of cinema theatres (Dibbets, 2006; Boter and Pafort-Overduin, 2009).

Central question

In the context of the research project CINEMAPS the authors aim to map cinema markets in the Netherlands and Flanders in the 1950s, 1960s and 1970s in a comparative study, combining a geospatial analysis of cinema density in both areas with data on pillarization, class and the organization and economics of industry. By projecting these data on historical maps in QGIS, the geographical distribution of different types of cinemas can be compared with patterns in cinema-going and local governmental policies in both countries. This multi-layered, international map functions as a heuristic tool to identify those areas that are interesting for further, in-depth analysis of the factors that explain the differences in market penetration. As such, the project will provide an answer to the core question of how the development of the cinema, as a specific cultural industry, interrelates with the social and cultural dimensions of modern public life in The Netherlands and Flanders, in particular pillarization, class and organization and economics of the industry. In this short paper we discuss our approach and method and present the first results.

Method

In the past decade, the use of GIS mapping technologies has proven a productive tool for analyzing the geo-spatial dimensions of cinema culture (Horak, 2016). The use of mapping is coherent with recent spatial orientations in film-historical scholarship, which “focuses on cinema as social experience, conditioned by factors such as transportation networks, ethnicity, and social group as well as cinema architecture, ticket prices, and the changing patterns of work and leisure.” (Hallam and Roberts, 2014: 20) Such a comprehensive, spatial approach can help to understand how cinema was experienced in the past and how cinema-going influenced the construction of social identity.

The use of GIS technology allows us to study the interrelation between cinema location and the social and economic dimensions of cinema culture at an unprecedented scale. For the comparative research on The Netherland and Flanders, we adopt a three-tiered approach. First, we map all the cinemas according to their typologies, distinguishing between permanent theatres, theatres with occasional screenings and travelling cinemas. Second, while acknowledging the fact that cartels such as the Netherlands Cinema Alliance (hereafter: NBB) extended a strong control over the cinema market (Van Oort, 2016), we also include local government data on the map, to acknowledge the influence of pillarized municipal policy on local cinema cultures. Finally, we map (expected) audiences in relation to their political ideology, religious denomination and income or class. In order to account for the limitations of geospatial technologies in explaining complex human cultural interaction (Verhoeven et al., 2009), in the next phase of our research we will use these maps to identify case studies for further, in-depth analysis.

Mapping cinema density in its socio-economic context – first results

The datasets used (census data, data on religious denomination, local election results, the Cinema Context and Verlichte Stad cinema databases) partly had to be digitized, and all of them had to be harmonized. The harmonization processes consists of equaling the granularity of comparison both nationally and internationally (e.g., comparing municipalities and cantons), the classification of categories (typologies of cinemas, political parties, and religious denomination), and solving discrepancies in periodization between the different datasets. For some datasets harmonization models are available (e.g., for religious orientation), for others they need to be created. The data on cinema locations in Flanders are currently being georeferenced and combined with the harmonized census data and other social datasets.

Since the Dutch data were most complete, in the first phase of the project we focused on mapping cinema density in the Netherlands. The map in Figure 1 shows the geographical distribution of Dutch commercial cinemas in relation to the expected cinema attendance. In general, we can conclude that the distribution of cinemas correlates with the level of expected cinema attendance: many permanent theatres in areas with a high expected attendance, mostly travelling cinemas in areas with a low expected attendance. This is not the case for the area in the middle and East of the country (cities of Apeldoorn and Enschede), which couples high attendance to low cinema density. This invites further research into the particularities of those areas.

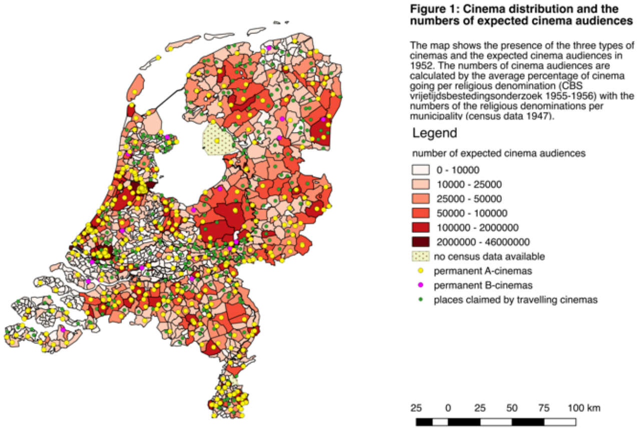


Figure 1: Cinema distribution and the numbers of expected cinema audiences

The map shows the presence of the three types of cinemas and the expected cinema audiences in 1952. The numbers of cinema audiences are calculated by the average percentage of cinema going per religious denomination (CBS vrijetijdsbestedingsonderzoek 1955-1956) with the numbers of the religious denominations per municipality (census data 1947).

Besides the influence of the organization and economics of the industry on the distribution of cinema theatres by the NBB, the local municipal policies were another major influence on cinemas and film screenings. Municipalities issued or refused permits for opening cinemas. Besides, they could prohibit certain film titles by arguing that they presented a threat to local order. Lastly, municipalities could impose (high) local taxes on cinema screenings. In short, this hitherto ignored data provides insights on the influence of pillarization on the distribution of cinemas.

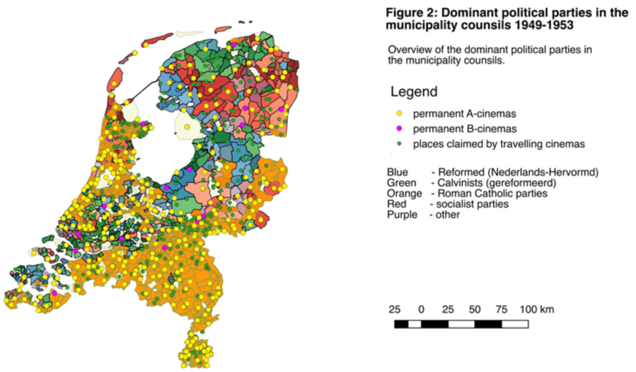


Figure 2: Dominant political parties in the municipality counsils 1949-1953

Overview of the dominant political parties in the municipality counsils.
Legend
● permanent A-cinemas
● permanent B-cinemas
● places claimed by travelling cinemas
Blue - Reformed (Nederlands-Hervormd)
Green - Calvinists (gereformeerd)
Orange - Roman Catholic parties
Red - socialist parties
Purple - other
25 0 25 50 75 100 km

The map in Figure 2 shows a correlation between protestant municipalities (known for discouraging cinema attendance) and low cinema density, whilst Catholic municipalities show a higher number of cinemas. This suggests that religious orientation did influence cinema market penetration, in spite of the supposed neutralizing role of the NBB.

Conclusions and future work

The first geospatial analysis of Dutch cinema culture yielded a number of preliminary conclusions. First, when identifying the data on Dutch cinema locations, it turned out that previous studies had ignored the large number of commercial cinema screenings in places frequented by travelling cinemas (310 in 1949). Although these cinema screenings constitute only a small percentage of the total cinema attendance (1%), the presence of these travelling cinemas does give rise to revisit the assumption that cinema density in the Netherlands was extremely low. Second, in some areas we see a high level of expected cinema-going, but very few cinema theatres - this can be researched further by combining the analysis with data on income and class of the expected audience, as well as further in-depth case study analysis. Third, the mapping of cinema density in relation to religious orientation questioned the assumption that pillarization was not relevant because of the neutralizing role of the NBB. These findings invite a more fine-grained study of local policies as well as film programming. In the next phase we will extend this study by comparing the Dutch cinema market to the Belgian one and use this comparative research to identify case studies for more in-depth research of local specificities of (not) going to the movies.

Bibliography

Biltereyst, D., and Meers, P. (2007). De verlichte stad: een geschiedenis van bioscopen, filmvertoningen en filmcultuur in Vlaanderen. Leuven: LannooCampus.

Boter, J., and Pafort-Overduin, C. (2009). "Compartementalisation and its influence on film distribution and exhibition in the Netherlands, 1934-1936." In Ross, M., Grauer, M. and Freisleben, B. (eds.), Digital Tools in Media Studies: Analysis and Research: An Overview. Bielefeld: Transcript, pp. 55–68.

Dibbets, K. (2006). "Het taboe van de Nederlandse filmcultuur: neutraal in een verzuild land." Tijdschrift Voor Mediageschiedenis, 9(2): 46–64.

Hallam, J., and Roberts, L. (eds) (2014). Locating the Moving Image: New Approaches to Film and Place. Indianapolis: Indiana University Press.

Horak, L. (2016). "Using digital maps to investigate cinema history." In Acland, C.R. and Hoyt, E. (eds.), The Arclight Guidebook to Media History and the Digital Humanities. Falmer: Reframe Books, pp. 65–102.

Oort, T. van (2016). "Industrial organization of film exhibitors in the low countries: comparing the Netherlands and Belgium, 1945–1960." Historical Journal of Film, Radio and Television, online first, published March 17: http://dx.doi.org/10.1080/01439685.2016.1157294 (accessed 1 September 2016).

Thissen, J. and Velden, A. van der (2009). "Klasse als factor in de Nederlandse filmgeschiedenis." Tijdschrift voor Mediageschiedenis, 12(1): 50-72.

Verhoeven, D., Bowles, K. and Arrowsmith, C. (2009) "Mapping the movies: reflections on the use of geospatial technologies for historical cinema audience research." In Ross, M., Grauer, M. and Freisleben, B. (eds.), Digital Tools in Media Studies: Analysis and Research: An Overview. Bielefeld: Transcript, pp. 69–82.

Text Mining the History of Information Politics Through Thousands of Swedish Governmental Official Reports

Fredrik Norén
fredrik.noren@umu.se
Umeå University, Sweden

Roger Mähler
roger.mahler@umu.se
Humlab, Umeå University, Sweden

How did "information", a concept and a keyword that we take for granted in our modern vocabulary, emerge into the official language? In a previous article for a special issue on digital methods in the journal *Nordicom-Information*, I analyzed the "voice" of the Swedish state and the concept of information in the twentieth century by topic modeling the collection of Swedish Governmental Official Reports (8 000 reports, 1922–). The scope and the long time-span of the corpora makes it an internationally unique source, especially when it comes to study the emergence of interests and attitudes of a single state through time. My results revealed that a topic of information barely existed before the 1960s, then dramatically exploded in frequency, and became a dominant part and parcel in the governmental discourse. In this presentation, I will show how, for example, the rise of an information topic should be understood in relation to a larger, and changing discursive context of Swedish politics.

The findings in this talk result from examining *co-occurring* topics, a less common approach within the practical use of probabilistic topic modeling. A method was used to challenge the previous historiography of information politics by situating an information topic within a larger cluster of topics, so called meta-topics, in all governmental reports from the 1960s, 1970s and 1980s. Studying co-occurring topics is important in order to better understand the nature and quality of probabilistic topic models, as well as the source material that topics represent (e.g. Kaufman 2016).

In collaboration with Humlab, the digital humanities center at Umeå University, and software developer Roger Mähler, topic modeling with Latent Dirichlet Allocation (LDA) was utilized to identify local topics (based on chunks of text) in the reports. Each topic was assigned an average weight based on all the local weights in the entire report. In addition, a co-occurrence value was also computed for all topic-pairs found in the entire corpora. The topic modeling was done using Mallet, and the rest of the calculations, using the output from Mallet, were implemented in Python scripts. The resulting data were visualized and explored as networks using the Gephi software.

This method was expected to give insight into three aspects: (1) the number and diversity of reports in which the information topic over time occurred in, (2) to discover and visualize larger cluster of topics to give a topic of information a position, and hence a context, in those clusters, and (3) enrich the analysis by combining distant and close readings. Thus, the presentation emphasizes the need for switching between a quantitative and qualitative approach, and argues that the findings can be used as a starting point for analyzing the concept of information and its contexts in ways that have not and could not have been done before.

The collection of reports was recently digitized by the National Library of Sweden (2015). The system of commission of inquiry constitutes a cornerstone of the Swedish governmental system. Before the government sends a bill to the parliament, they often appoint a commission to investigate different alternatives of the bill, and the result is a governmental report. Due to the long time-span, the reports cover a wide range of political issues. And, it is the rich diversity of political issues that enables us to view the collection as "the voice" of the Swedish state, similar to what Franco Moretti and Dominique Pestre did when they explored the language used in the World Bank Reports (Moretti and Pestre 2015). Today, incalculable amounts of texts (such as this collection of reports) are not only available online but are also searchable, even down to individual words. This creates a challenge for the humanities to potentially re-rewrite parts of history. That is to say, the digital humanities are now tasked with investigating the ways that changes in language, across and within millions of documents, can be linked to– and thus create new understanding of– developments in society (Jockers 2013).

The Swedish Governmental Official Report series are available for public access at *The Riksdag's open data* website (data.riksdagen.se), and part-of-speech tagged versions are available at Språkbanken (spraakbanken.gu.se) as downloadable XML-files. This study extracted the word stem of all nouns from the reports of the 1960s, 1970s and 1980s, which is sufficient for the study of themes in a text. Each report was split into chunks of 1 000 nouns each (following the work of Jockers 2013). Mallet was then used to compute three distinct LDA topic models, one for each decade, and each consisting of 500 topics. Different numbers of topics were tried out before we settled on 500, based on a breadth variety of what was perceived as concrete political issues (topics). A manual review of the generated topics showed that each decade had been assigned a distinctive topic of a general theme of information. The computed average topic weights for each report, based on weights in each chunk, were used to visualize a network of all reports, and their most dominant topics, with weaker report-to-topic links filtered out based on a configurable threshold. Our pre-study showed that LDA topic modeling often computes a very dominant topic with a weight of over 60%, while the weight of the following topics dropped significantly. Hence, a generous threshold of 0.01 was proved to capture both a discursive core as well as its periphery.

The method used in this study showed three things. Firstly, a distinctive theme of information evolved over time in the official language of the Swedish state bureaucracy. By highlighting reports, in which the information topic was dominant, it was clear that the number of reports increased by a fourfold from the 1960s to the 1980s. Also, by categorizing the reports by state minister of origin, it was revealed that the information topic was associated with growing numbers of political issues over time.

Secondly, the three datasets, one for each decade, were imported, separately, to Gephi. The Force Atlas and Modularity Class algorithms was used to sort topics and reports into larger thematic clusters, or meta-topics. For each topic, the top keywords belonging to a cluster were manually examined as a way to identify the common theme of the topics and the cluster itself. This also allowed for examination of the connections found between the topic of information and other topics and reports.

Thirdly, a distant perspective does not neglect the possibility of adding close reading to the analysis. On the contrary, by zooming in on the topic cluster, it became clear that the reports in which the information topic had a strong presence were often those we least expected. The actual text of reports on foreign policy, research and innovation (among others) presented concrete insights and illustrations of how to understand and synthesize the connection between "information" and various political issues, while also demonstrating the necessity of the concept of information when dealing with various problems and challenges in the post-war society. Hence, the method proved to be a platform for a dynamic interaction between the very close and the very distant that, ideally, could help to strengthen both the quantitative and qualitative perspective as well as the end-result.

Bibliography

Kaufman, M. (2016). "'Everything on Paper Will Be Used Against Me': Quantifying Kissinger. Text Analysis, Visualization and Historical Interpretation of the National Security Archive's Kissinger Correspondence." http://blog.quantifyingkissinger.com/.

Jockers, M.L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

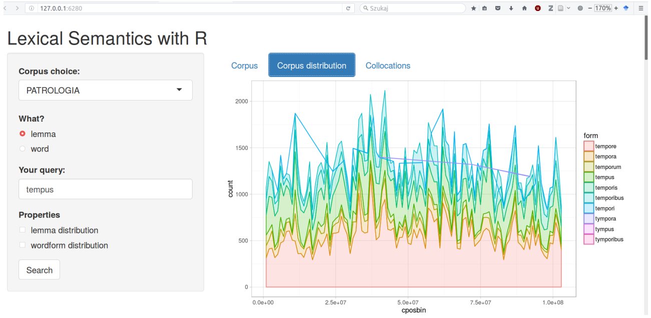
Moretti, F. and Pestre, D. (2015). "Bankspeak: The Language of World Bank Reports, 1946–2012". http://litlab.stanford.edu/LiteraryLabPamphlet9.pdf

Voces. An R-based Dashboard for Lexical Semantics

Krzysztof Nowak
krzysztof.nowak@ijp-pan.krakow.pl
Institute of Polish Language
Polish Academy of Sciences Krakow, Poland

Introduction

Voces (from Lat. *vox* 'voice', 'word') is an analysis and visualisation dashboard for corpus-based research in lexical semantics. Currently developed as a Shiny application communicating with R session running in the background, *Voces* provides users with possibly exhaustive account of how selected Latin word is distributed across the corpus and what can be told about its meaning. The application is built around a corpus which currently consists of ca. 200M words from texts dating from the Classical era (1 BCE) to the Middle Ages (14th CE). Although *Voces* was originally conceived as a tool of historical semantics research, the application - due to its modular design - may be modified and the code basis can be re-used in new research contexts.



Lexical Semantics with R

Corpus choice: PATROLOGIA

What?
● lemma
○ word

Your query: tempus

Properties
□ lemma distribution
□ wordform distribution

Fig. 1: Voces. User Interface: Word Form Distribution (tempus 'time')

Information computed on a basis of a CWB-indexed corpus is presented to a user through a single-page interface composed of separate widgets arranged in a clear grid layout. Each widget is responsible for displaying in textual or graphical form a clear-cut property of word's distribution or meaning. A heavy use of data visualisation techniques renders *Voces* a convenient tool for exploratory analysis of textual corpora, but the grid layout is also reflection of modular architecture of the application. Each widget is implemented as a separate function which can be extended and adopted by researchers with even limited R programming skills.

Use scenarios

A typical use scenario is triggered when the user specifies a lemma to be looked up. If the search fails, a list of lemmas to choose from is provided. In case of success, neatly separated sections of the dashboard are populated with widgets, each of which corresponds to one sense or distributional property of the word under scrutiny.

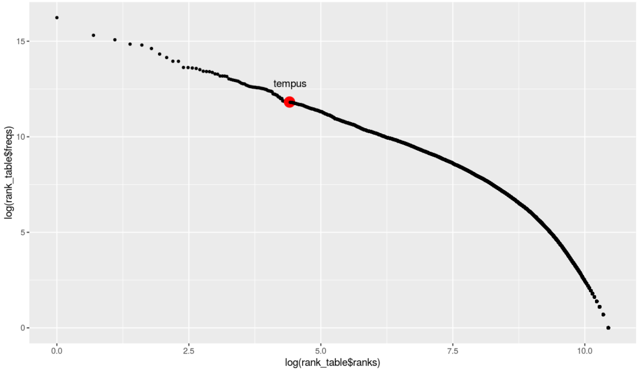


Fig. 2: Voces. User Interface: Frequency Spectrum Plot (Voces. User Interface (tempus 'time')

Word's frequency is summarised as a number of occurrences in the corpus (both raw and p.m.w. counts) and displayed as a highlighted point on a frequency spectrum plot (Baayen 2001). A barplot is provided for investigating change of frequency in subsequent corpus sections. Study of language variation is enabled through widgets presenting word's frequency as a function of such variables as author, work, genre, and – most importantly – time. Users are, therefore, provided with a list of authors who use the word most frequently or a word cloud summarising terms to be found in the titles of works with a particularly frequent use of the word under scrutiny. Genre variation is presented in form of a pie chart, while diachronic dimension - through a bar plot of frequency counts in partitions of the corpus. Diatopic variation study is still to be implemented.

A word's meaning potential can be investigated by means of a set of widgets presenting its contextual properties. The most frequent co-occurrences are enumerated on a simple count list which may be further analysed according to period and genre criteria. A Distributional Semantics Model (Baroni and Lenci 2010) is built from the corpus in order to enable simple meaning computation. Evert's (2014) *wordspace* package and a set of Alain Guerreau's scripts is employed in order to cluster co-occurrences. Similar terms of a looked up word are also computed and then presented in both textual and graphical form.

Users are supported in data and visualisation interpretation through hints which accompany every widget. Their role is to explain not only what the data can mean, but also how the figures were computed, how one can interpret the geometrical properties of a plot, and so on. This, along with the availability of data sets, code snippets, and reports generated on the fly, is what makes *Voces* a tool of reproductive research.

Architecture

Voces was built as a Shiny application (Chang et al. 2016). Its development was greatly facilitated by the availability of a decent documentation and community support (both particularly useful when dealing with framework's complex reactivity model). It turned out soon, however, that it may not be the best choice for web application which has to combine heterogeneous data and non-R code as well. Hence, other solutions are being tested at the moment, those in particular which would provide, for example, more flexible integration of external APIs. The most promising seems to be OpenCPU (Ooms 2014), an application which exposes R session through a RESTful API. This approach allows any application written in some of the less or more popular web development frameworks to easily communicate with an R server instance.

As for the architecture, *Voces* depends on a CQP server instance running in the background which requires corpora to be indexed with the CWB. Communication of the R server with the CWB is assured through the rcqp package (Desgraupes and Loiseau 2012) which offers a set of useful functions providing access to both positional (token-level) and structural (document-level) attributes. Unfortunately, development of this very helpful tool seems to be less active recently and thus *Voces* will soon accept also tabular data as input.

Previous research

Nowadays, corpus linguists may chose from a vast array of free, open source and stable corpus query systems (CQS) which not only allow for efficient indexing of large corpora, but also provide a user-friendly concordance interface and offer out-of-the-box a set of such essential functionalities as collocation lists, simple corpus statistics etc. Both web (CQPweb, NoSketchEngine *etc.*) and desktop applications (TXM *etc.*) are also usually equipped with a less or more intuitive corpus management interface. *Voces*, a dashboard for vocabulary research, is not yet another CQS and has no intention to supersede well-established tools which cannot be easily combated in terms of either robustness or speed. Quite the contrary, the application communicates with the CWB engine and adapts some of the design choices and features of the popular CQS, while hopefully does not inherit their drawbacks.

Unlike the case of the well-known CQS, more emphasis has been put on quick access to multifaceted information rather than on close analysis of occurrences. *Voces* does not attempt, then, to implement some of the features which are traditionally considered an important part of the corpus analytical toolbox, such as concordance sampling, sorting etc. Undoubtedly, the strength of popular CQS lies in their wide applicability: by default, they do not preclude any research scenario. Although agnostic of linguistic theory, *Voces* was originally built for more specific purposes and focuses on semantic properties of the word and its distribution.

What is believed to be one of the main advantages of the present application is that - thanks to its modular architecture - it can be easily extended or adopted by a researcher with even moderate programming skills. In that *Voces* attempts to fill the gap that exists between, on the one hand, fully-blown CQS, which are normally quite conservative when it comes to adding new features, and, on the other hand, single-purpose research workflows built *ad hoc* by researchers. What also distinguishes *Voces* from other CQS is its emphasis on helping users to interpret data. A system of visual and textual hints keeps a researcher informed about where does the data come from, how have they been computed *etc*.

The grid layout is well-known from analytical environment and is especially popular in finances or engineering (Few 2013); in humanities it was adopted, among others, in the Voyant Tools project. It offers a quick insight into otherwise dispersed data and a coherent account of word's properties.

Further research

Voces is currently in an early stage of development. The work focuses on adding new functionalities and plotting types which may sometimes affect application's efficiency. Future work will focus on: (1) optimising user experience; (2) implementing tools for (a) comparative (ie. two-lemma) research and (b) tracking language change; (3) better processing user input (multi-word search).

Bibliography

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Baroni, M., and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36 (4): 673–721.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2016). *Shiny: Web Application Framework for R.* https://CRAN.R-project.org/package=shiny.

Desgraupes, B., and Loiseau, S. (2012). *Rcqp: Interface to the Corpus Query Protocol.* http://CRAN.R-project.org/package=rcqp.

Evert, S. (2014). Distributional Semantics in R with the Wordspace Package. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 110–114. Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Few, S. (2013). *Information Dashboard Design: Displaying Data for at-a-Glance Monitoring*. Burlingame, CA: Analytics Press.

Nowak, K., and Bon, B. (2015). *Medialatinitas.eu*. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin. In *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 Conference*, edited by Iztok Kosem, Miloš Jakubíček, Jelena Kallas, and Simon Krek, 152–69. Ljubljana-Brighton: Trojina, Institute for Applied Slovene Studies - Lexical Computing Ltd.

Ooms, J. (2014). The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. *ArXiv E-Prints*, June.

A “Wind of Change”: Shaping Public Opinion of the “Arab Spring” Using Metaphors

Alexandra Núñez
nunez@linglit.tu-darmstadt.de
Institut für Sprach- und Literaturwissenschaft
TU Darmstadt, Germany

Malte Gerloff
gerloff@kdsl.informatik.tu-darmstadt.de
Institut für Philosophie
TU Darmstadt, Germany

Erik-Lân Do Dinh
dodinh@kdsl.informatik.tu-darmstadt.de
UKP Lab, TU Darmstadt, Germany

Andrea Rapp
rapp@linglit.tu-darmstadt.de
Institut für Sprach- und Literaturwissenschaft
TU Darmstadt, Germany

Petra Gehring
gehring@phil.tu-darmstadt.de
Institut für Philosophie, TU Darmstadt, Germany

Iryna Gurevych
gurevych@cs.tu-darmstadt.de
UKP Lab, TU Darmstadt, Germany

Motivation

How does mass media affect the way we think about controversial topics such as the “Arab Spring”? What persuasive role do metaphors play especially in opinion pieces?

During the events of the years 2010–2011 in the Middle East & North Africa region a new discourse was established in the German media; immediately these events were assessed as a “wave” of democratization and liberation, and have been metaphorically labeled “Arab Spring”. Metaphors were frequently used to categorize and understand these events (Möller, 2014; Núñez, 2014).

Given the premise that mass media organizes (Couldry, 2010) and shapes social reality (Luhmann, 1996), we analyze how the Arab Spring is categorized and assessed using metaphorical constructions in newspaper opinion pieces. We show ways in which particularly the use of metaphors reveals how the media tried to achieve acceptance for the events based on our cultural models (Quinn and Holland, 1987), which are grounded on our western knowledge.

According to the *Conceptual Metaphor Theory* (Lakoff and Johnson, 1980) metaphors are ubiquitous and exhibit a binary source-target domain structure. The knowledge that we choose to function as a source domain illustrates which conventionalized, overt or tacit knowledge we require to understand new or abstract domains (target domains) in terms of our cultural imprints. Metaphors are instantiated on the text surface and give us clues toward our knowledge basis. Thus, the required knowledge can be described in terms of ubiquitous metaphorical patterns that function as semantic “anchors” in texts, and in terms of conceptual knowledge clusters that function as an intertextual semantic knowledge structure.

As such, we constructed a pipeline that automatically detects metaphors appearing within certain grammatical constructions, before clustering them by presumed source and target domains. The results give us insights into how the Arab Spring is metaphorically structured by semantic clusters in opinion pieces.

Corpus and annotation

Our corpus consists of 300 opinion pieces (Ramge and Schuster, 2001) from five German newspapers, *Frankfurter Rundschau, Die Zeit, Der Spiegel, taz,* and *Die Welt*, which have been written between December 2010 and November 2011 and cover the Arab Spring.

In nine of these opinion pieces, two of this abstract’s authors annotated following grammatical constructions whether they constitute metaphors: adjective-noun (AN) pairs (e.g. “Tunisian spark”), and genitive constructions (GEN) (e.g. “torch of freedom”). Due to their interrelated components they provide a good insight into the structural systematicity of metaphorical mappings (source domain → target domain). The difficulty of the task is reflected in a low inter-annotator agreement (0.45 Krippendorff's α).

Common sources of annotation disagreement included heavily conventionalized metaphors such as "social network", personifications like "self-consciousness of a generation", and metaphors that need a larger context to function. As gold standard for further training and evaluation we only use the agreed upon annotations ("annotated", Table 1).

	Sentences	AN constructions	AN metaphors	GEN constructions	GEN metaphors
annotated	538	968	116 (12%)	102	29 (28%)
complete	11402	19573	-	2758	-

Table 1: Constructions and metaphors in the corpus.

Technical realization

To examine our questions quantitatively, we contrast two approaches to automatically detect metaphors, namely random forests and multilayer-perceptron. The extracted metaphors are subsequently clustered (Figure 1). To extract AN and GEN constructions we first perform automatic preprocessing, including part-of-speech tagging and dependency parsing.

The random forests approach of Tsvetkov (2014) firmly roots in conceptual metaphor theory, mainly employing features extracted from manually crafted resources such as an abstractness wordlist and supersenses, to classify adjective-noun and subject-verb-object constructions. For use on other languages than English, a bilingual dictionary is required. We manually expand an existing dictionary to cover our corpus, and extend their system to classify GEN metaphors.

The described feature-rich approach will be compared – with regards to what (kind of) metaphors can be found – to the shallow neural network approach by Do Dinh and Gurevych (2016), which does not presuppose any specific metaphor theory. It thus does not make use of external features, but rather learns exclusively from given annotations and their context. Preliminary experiments show that more training data is needed for this bottom-up approach.

To gain further insight into usage of metaphor in our corpus, we cluster the automatically found metaphors – resp. their components – into coarse-grained semantic fields. While there are works using a theory-supported top-down approach, e.g. using source domain lists (Gordon et al., 2015), we employ a rather unsupervised approach, without preselecting the number of clusters or manually fixing cluster centers (similar to Shutova et al. (2010), who use spectral clustering for metaphor detection). To that end, we employ Affinity Propagation (Frey and Dueck, 2007), which we supply with cosine similarities between pre-trained word embeddings (Reimers et al., 2014) of the metaphor components.

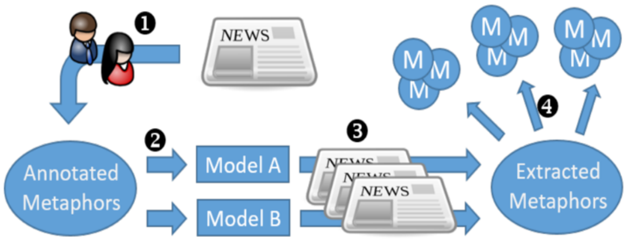


Figure 1: Few newspaper opinion pieces are annotated (1) and the obtained metaphors are used to learn different models A and B (2) (Tsvetkov, 2014; Do Dinh & Gurevych, 2016). These models are then applied to more articles to extract metaphors (3), which are subsequently clustered (4).

Experiments and discussion

We use cross-validation for the intrinsic evaluation of the metaphor detection part. For GEN metaphors, the tested system achieves 0.63 precision, 0.25 recall, and 0.35 F_1-score, with similar performance for AN metaphors. While these results seem low, the actual output of the system when trained on all annotated instances looks promising, and the precision is improved by filtering based on named entities.

Kopf: *Schwertern* des Islams, *Köpfen* des verhassten Regimes, *Handlanger* des Regimes, *Schlägern* des Regimes, *Arm* des alten Regimes, *Zähne* eines Kamms, *Kugeln* des Regimes, *Brust* des Leblosen, *Gesicht* des Exzesses, *Gesicht* der EZB, *Gesicht* des Arabischen Frühlings, *Gesicht* arabischer Demokratien, *Gesichter* der Demonstranten, *Gesichter* der Jasmin-Revolution, *Gesicht* der Revolution, *Gesicht* der Muslimbruderschaft, *Gesicht* der ägyptischen Revolution, *Gesicht* des Landes, *Gesicht* der tunesischen Revolution, *Gesicht* der USA, *Gesicht* ihres Sohnes, *Gesichter* der Vermissten, *Fortsatz* des alten Regimes, *Wand* der Angst, *Loch* der Diktatur, *Schweine* des Regimes, *Augen* vieler Araber, *Hände* des israelfeindlichen Regimes, *Hand* der Sozialisten, *Hände* des Obersten Kommandorates, *Händen* der Börse

Figure 2: GEN metaphors clustered by first noun, with center "Kopf", hinting at conceptual metaphor POLITICAL SYSTEMS ARE BODIES

The clustering creates an impression of which knowledge (source domain) is required for abstract concepts (target domains), and how abstract concepts are "perspectivized" in the corpus, while also giving an overview of occurring intertextual metaphors. Although the cluster assignment and the metaphor detection leave room for improvement (e.g. Figure 2: "face of her son"), the clusters already reveal the systematicity and constraints of metaphorical mappings. Thus, they point to strategies of newspapers that come along with the choice of the (conceptual) source domain.

In Figure 2, bodily parts such as *face*, *head*, *hand* are used as source domains and mapped to political systems or processes (e.g. *regime, democracy, revolution*). This mapping draws on a long tradition in political and philosophical history (Musolff, 2004): *head* and *face* play a central role in our culture – comparing political processes with *faces* or *heads* conceptualize them as human beings. In this cluster the construction *face of* indicates that the events are important, thus construed as worthy to support.

Those prototypical examples for ontological metaphors also support the premise of embodied cognition (Johnson, 1987; Rohrer, 2010).

Sturm: *Schlussphase* des alten Regimes, *Sturm* des arabischen Umbruchs, *Sturm* der Moleküle, *Sturm* der Entrüstung, *Inseln* der Diktatur, *Insel* der Stabilität, *Wind* der Freiheit, *Wind* der Demokratie, *Wind* der Revolution, *Wind* des Wandels

Figure 3: GEN metaphors clustered by first noun, with center "Sturm", hinting at conceptual metaphor POLITICAL SYSTEMS/PROCESSES/VALUES ARE NATURAL ELEMENTS

The positive properties and the movement character of natural elements such as *wind* and *storm* are mapped to the abstract (political) nouns *freedom*, *revolution,* or *change*. They receive a deontic (Hermanns, 1994) character, whereas *dictatorship* is conceptualized in terms of *island* which stands for inertia and stability (Figure 3). These examples already show how the chosen metaphors shape dualistic tendencies by categorizing the events on one hand as a dynamic movement (*wind, storm*) that has to be supported by western democracies, or on the other hand pleading for stability (*island*), thus implicitly supporting dictatorship.

The analyzed clusters and metaphorical conceptualizations indicate a network of source domains that function as key concepts which structure the discourse of the Arab Spring, an assumption we will focus on in future work.

Conclusion and future work

Our study indicates that metaphorical constructions are important in media because of their ubiquitous use in opinion pieces. The generic extracted source domains already suggest that a specific network of knowledge is used in media to highlight certain political aspects of the Arab Spring. Furthermore, they illustrate how contents are emotionalized and ideologized during these events by metaphors and their framing effects. Usage of natural elements or body parts reduces complexity and conceptualizes the events as an organic development, in short: the Arab states become western democratic states. Thus contributing to the extension of western ideology, metaphors impart implicit cultural values. Combining our cognitive and discourse analytical questions we can summarize that the used "bottom-up" clustering is very helpful to get an explorative impression of the "intertextual consistencies" (Verschuren, 2012: 179) of chosen metaphors. They are good textual "anchors" and starting points to investigate the widespread metaphorical use, and thus knowledge domains, in corpora.

With regard to the state of the art, corpus-based methodologies within the Digital Humanities community will benefit from our research by gaining the possibility to automatically compare thematic corpora by using the relationship of their metaphors to the common main cluster as a metric, therefore obtaining a new way to analyze the conceptual network being used. Our approach can help to facilitate corpus studies, e.g. by analyzing other discourse segments which deal with the implicit construction of identity and alterity within opinion pieces by using metaphors.

In our presentation we will highlight the results and give a structured impression of the mappings and the implications of the used metaphors in our corpus and present in short our methodological basis.

In future studies we will compare the conceptualization strategies of the Arab Spring and "Refugee Crisis" in German media, since we assume that the same metaphors and the same (metaphorical) interpretation patterns occur. Further, we plan to investigate another theory of metaphor which is based on Black (1954, 1977) and Gehring (2010). The latter model is strongly interweaved with current discussions about "Begriff" (Müller-Meiningen and Schmieder, 2016; Gehring, 2005, 2010) and discusses its ideological implication(s). Furthermore, the emphasis is placed on the function of metaphors as an epistemological tool by investigating, amongst others, the evolution of ideas and cultural values, e.g. in the historical text collection "Natur&Staat" (1903-11).

Bibliography

Black, M. (1977). More about Metaphor. Dialectica, 31: 431–457.

Black, M. (1954). Metaphor. Proceedings of the Aristotelian Society. New Series, 55: 273–294.

Couldry, N. (2010). Media discourse and the naturalization of categories. In Wodak, R. and Koller, V. (eds), Handbook of communication in the Public Sphere. Berlin / New York: Walter de Gruyter, pp. 67–88.

Dirven, R., Polzenhagen, F., and Wolf, H.-G. (2010). Cognitive Linguistics, Ideology, and Critical Discourse Analysis. In Geeraerts, D. and Cuyckens, H. (eds), The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford University Press, pp. 1223–1240.

Do Dinh, E., and Gurevych, I. (2016). Token-Level Metaphor Detection using Neural Networks. In Proceedings of the Fourth Workshop on Metaphor in NLP. San Diego, CA, USA: Association for Computational Linguistics, pp. 28–33.

Frey, B. J., and Dueck, D. (2007). Clustering by Passing Messages Between Data Points. Science, 315(5814): 972–976.

Gehring, P. (2005). Vom Begriff zur Metapher. Elemente einer Methode der historischen Metaphernforschung. In Abel, G. (ed), Kreativität. Kolloquiumsbeiträge des XX. Kongresses der Allgemeinen Gesellschaft für Philosophie in Deutschland. Hamburg: Meiner, pp. 800–815.

Gehring, P. (2009). Das Bild vom Sprachbild. Die Metapher und das Visuelle. In Danneberg, L., Spoerhase, C., and Werle, D. (eds), Begriffe, Metaphern und Imaginationen in Philosophie und Wissenschaftsgeschichte. Wiesbaden: Harrassowitz, pp. 81–101.

Gehring, P. (2010). Erkenntnis durch Metapher? Methodologische Bemerkungen zur Metaphernforschung. In Junge, M. (ed), Metaphern in Wissenskulturen. Wiesbaden: Verlag für Sozialwissenschaften, pp. 203–220.

Gehring, P., and Gurevych, I. (2014). Suchen als Methode? Zu einigen Problemen digitaler Metapherndetektion. Journal Phänomenologie, 41: 99–110.

Gordon, J., Hobbs, J. R., May, J., Mohler, M., Morbini, F., Rink, B., Tomlinson, M., Wertheim, S. (2015). A Corpus of Rich Metaphor Annotation. In Proceedings of the Third Workshop on Metaphor in NLP. Denver, CO, USA: Association for Computational Linguistics, pp. 56–66.

Hermanns, F. (1994). Schlüssel-, Schlag- und Fahnenwörter; zu Begrifflichkeit und Theorie der lexikalischen "politischen Semantik". Erste Fassung eines Überblicksartikels zum Forschungsstand in Sachen Schlüsselwort- und Schlagworttheorie und -forschung für den Ergebnisband de. Heidelberg.

Johnson, M. (1987). The body in the mind. The bodily basis of meaning, imagination, and reason. Chicago: Chicago University Press.

Lakoff, G. (2006). Conceptual Metaphor. The contemporary theory of metaphor [1993]. In Geeraerts, D. (ed), Cognitive Linguistics: Basic Readings. Berlin: Mouton de Gruyter, pp. 185–238.

Lakoff, G., and Johnson, M. (1980). Metaphors we live by. Chicago: Chicago University Press.

Möller, N. (2014). Cognitive Metaphor and the "Arab spring." In Polzenhagen, F., Kleinke, S., Kövecses, Z., and Vogelbacher, S. (eds), Cognitive Explorations into Metaphor and Metonymy. Frankfurt (Main): Peter Lang, pp. 133–148.

Musolff, A. (2003). The heart of the European body politic: British and German perspectives on Europe's central organ. Journal of multilingual and multicultural development., 25: 437–452.

Müller-Meiningen, E., and Schmieder, F. (2016). Begriffsgeschichte und historische Semantik. Berlin: Suhrkamp Verlag.

Luhmann, N. (2009). Die Realität der Massenmedien. Wiesbaden: Verlag für Sozialwissenschaften.

Núñez, A. (2014). Wenn das "Embodiment" politisch wird: Das Image-Schema PATH und seine Realisierung im Mediendiskurs zum „Arabischen Frühling". In Polzenhagen, F., Kleinke, S., Kövecses, Z., and Vogelbacher, S. (eds), Cognitive Explorations into Metaphor and Metonymy. Frankfurt (Main): Peter Lang: pp. 149–164.

Quinn, N., and Holland, D. (1987). Culture and Cognition. In Holland, D. and Quinn, N. (eds), Cultural models in language and thought. Cambridge: Cambridge University Press, pp. 3–42.

Ramge, H., and Schuster, B.-M. (2001). Kommunikative Funktionen des Zeitungskommentars. In Leonhard, J.-F., Ludwig, H.-W, Schwarze, D., and Straßner, E. (eds), Medienwissenschaft. Ein Handbuch zur Entwicklung der Medien und Kommunikationsformen. Berlin / New York: Mouton de Gruyter, pp. 1702–1712.

Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., and Gurevych, I. (2014). GermEval-2014: Nested Named Entity Recognition with Neural Networks. In Workshop Proceedings of the 12th Edition of the KONVENS Conference. Hildesheim, Germany: Universitätsverlag Hildesheim, pp. 117–120.

Rohrer, T. (2010). Embodiment and experientialism (2010). In Geeraerts, D. and Cuyckens, H. (eds), The Oxford Handbook of Cognitive Linguistics. Oxford: Oxford University Press, pp. 25–47.

Shutova, E., Sun, L., and Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In Proceedings of the 23rd International Conference on Computational Linguistics. Shanghai, China: Association for Computational Linguistics, pp. 1002–1010.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor Detection with Cross-Lingual Model Transfer. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 248–258.

Verschuren, J. (2012). Ideology in Language Use. Pragmatic Guidelines for Empirical Research. Cambridge: Cambridge University Press.

Ziegler, H.E., Conrad, J., Haeckel, E. (eds) (1903–1911). Natur und Staat, Beiträge zur naturwissenschaftlichen Gesellschaftslehre. Jena: Gustav Fischer Verlag.

All the Things You Are: Accessing An Enriched Musicological Prosopography Through *JazzCats*

Terhi Nurmikko-Fuller
terhi.nurmikko-fuller@anu.edu.au
Australia National University, Australia

Daniel Bangert
d.bangert@unsw.edu.au
UNSW Australia, Sydney, Australia

Alfie Abdul-Rahman
alfie.abdulrahman@oerc.ox.ac.uk
University of Oxford, United Kingdom

Introduction

JazzCats (Jazz Collection of Aggregated Triples) is a prototype project which uses Linked Open Data (LOD) to support musicological, historical, and prosopographical analyses. It has increased access to (and the openness of) data published online through a twofold process: firstly, information hitherto unavailable to users has been shared and incorporated into the project, and secondly, data previously locked in non-Open types (e.g. PDF) has been published in a machine-readable format, increasing discoverability in the context of the wider Web. Connections between datasets that could only be identified through a human user engaging separately with each existing project have now been made explicit, and the resulting aggregated data is queryable from a single user-interface (UI).

Three projects contribute to *JazzCats*: a social network connecting musicians through various types of relationships is provided by LinkedJazz (Pattuelli, 2016); details of solos within performances (including pitch, key, and chord changes) are available from WJazzD (Pfleiderer et al., 2016); and Body&Soul (Bowen, 2013) is a discography of over 200 recordings. These complementary data contain instance-level overlaps for recordings and musicians. Bringing these resources together has enabled a new type of research question, possible only through using criteria from one dataset to inform and hone results from another.

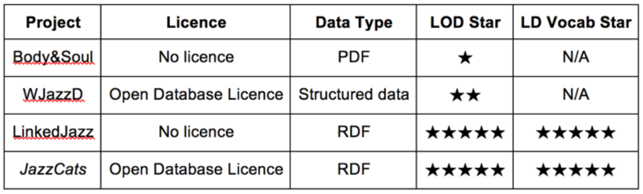
Limitations of existing data publication methods

The sub-projects at the heart of *JazzCats* engage with the 5 Star Standard of LOD (Berners-Lee, 2006) to different extents (see Table 1). The data for Body&Soul has been published online as a PDF, making it an ideal example of 1

Star categorization. WJazzD allows users to download both the database and the software (2 Star). LinkedJazz provides two separate data-dumps of RDF (with an additional, earlier set of triples also available), containing both dereferenced URIs and those which point to human-readable pages. We have categorized this project as 5 Star, because DBpedia resource URIs (related through **owl:sameAs** relationships to LinkedJazz resource URIs) are used (even if the retrieving of additional data from external sources is currently not possible via the LinkedJazz SPARQL endpoint). Publishing the information from the first two datasets with distinct HTTP URIs, connecting them to each other as well as to the RDF acquired from LinkedJazz, makes *JazzCats* 5 Star standard.

Conversion to Linked Data (LD) does not automatically ensure that information is more reusable or discoverable by data consumers on the Web (Bechhofer et al., 2013; Janowicz et al., 2014). Closed systems can benefit from LD, and whilst adherence to the LOD paradigm is an essential criterion for enabling reuse of any project's RDF by other data publishers, effective queries by a wider base of users can be restricted by idiosyncratic or project-unique vocabularies. To encourage good practices, Janowicz et al. (2014) have introduced a 5 Star rating, ranging from LD without vocabulary use (0 Star) to a vocabulary that is linked to by other vocabularies (5 Star). The term vocabulary is used in a broad sense to include all types such as schemata, and ontologies. We have categorized LinkedJazz as 5 Star because it links to other vocabularies and metadata about the vocabulary is available.

JazzCats makes extensive use of properties and classes from other vocabularies, including the Music Ontology (MO) (Raimond et al., 2007), the Event Ontology (Raimond and Abdallah, 2007), FOAF (Brickley and Miller, 2014), and SKOS (Miles et al., 2005). It is currently classified as a 5 Star since metadata about the *JazzCats* ontology is available in a dereferenceable and machine-readable form, but other vocabularies do not yet link to *JazzCats.*



Project	Licence	Data Type	LOD Star	LD Vocab Star
Body&Soul	No licence	PDF	★	N/A
WJazzD	Open Database Licence	Structured data	★★	N/A
LinkedJazz	No licence	RDF	★★★★★	★★★★★
JazzCats	Open Database Licence	RDF	★★★★★	★★★★★

Table 1: Evaluation of the JazzCats composite projects

Increasing accessibility through *JazzCats*

A previously unpublished CSV containing Body&Soul data was cleaned and enriched with additional information held in PDF files using OpenRefine (2013) to create a new, open access dataset (Bangert, 2016). An existing workflow (Nurmikko-Fuller et al., 2016) was then reproduced to map this tabular data into RDF using an Open-Source data integration tool (Web-Karma). This workflow relied on domain expert user-input to complete the ontological modeling and instance mapping stages within Web-Karma (University of Southern California, 2016). To support the future alignment and enrichment of this data with other musicological datasets, the underlying ontological structure extensively utilizes the properties and classes of the MO. The data structure has been documented on the website of the *JazzCats* project (Bangert et al., 2016).

Both the data and the software for WJazzD are available for download from the Jazzomat Research Project website. Although structured data, and in adherence with the 2 Star LOD publication criteria, information in this form is not accessible for machine-inferencing, and the clustered tables can be difficult for human users to navigate. The data was converted to RDF by repeating a second workflow as described in Nurmikko-Fuller et al. (2016) using the D2R server (Cyganiak and Bizer, 2012). This automated process produced clusters of triples based on database information categories (e.g. melody, beats, sections), which are mostly expressed through **xsd:strings** and **xsd:integers**. Mappings were made where applicable to connect these elements together using MO properties and classes. An overview of the ontological structure, and a detailed subsection illustrating the different properties are documented and defined on the *JazzCats* website.

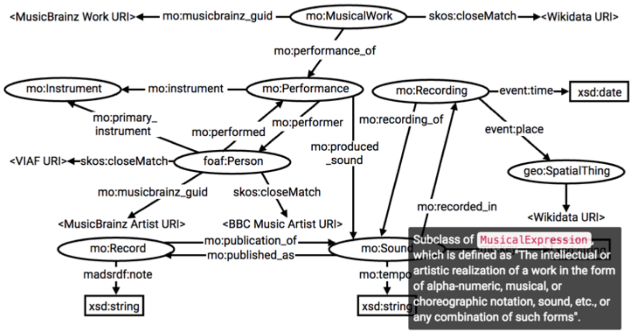
LinkedJazz provides two separate datasets for entities and the 12 different types of interpersonal (both professional and social) relationships between them. Adding these RDF-dumps to the *JazzCats* triplestore enables queries combining this prosopography with performance metadata derived from the other projects which make up the entirety of the *JazzCats* data.

Publishing these datasets as RDF using common vocabularies and ontologies known to have been utilized in other digital musicology projects increases their discoverability and the value. As data publishers, adhering to LOD standards allows us to further benefit from any additional future linkage. A conscious decision has been made at the onset of the *JazzCats* development process to publish RDF, ontologies, and raw data in Open and accessible formats, with appropriate licensing, to allow for the replication of our workflow, verification of our findings, and reuse of any or all of the composite parts of the project. *JazzCats* is made available under the Open Data Commons Open Database License (Miller et al., 2008). The aim throughout has been to produce and publish data that adheres to the FAIR data principles of being findable, accessible, interoperable and reusable (Wilkinson et al., 2016).

Evaluation of *JazzCats*

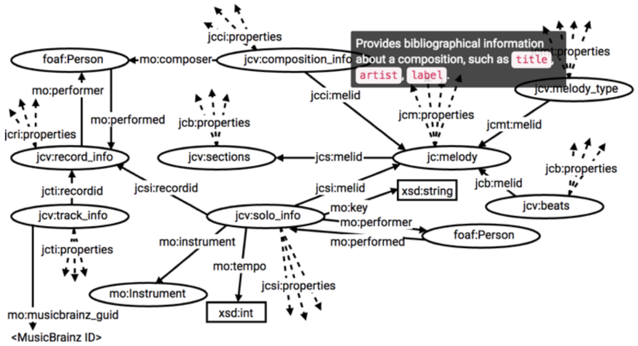
Although the data in *JazzCats* adheres to a 5 Star standard of LOD for accessibility and openness, the current UI presents a barrier to access for human users. At present, the project RDF (RDF Core Working Group, 2014) is contained within an instance of the Open-Source version of the Virtuoso (OpenLink Software, 2016) triplestore, and is only queryable using the default SPARQL endpoint, accessible from the *JazzCats* website. Example queries demonstrate how results can be generated by drawing from all three datasets simultaneously (Nurmikko-Fuller, 2016), but engagement with the data beyond the parameters set by these existing samples requires the user to have the necessary skills to construct new SPARQL queries (W3C RDF Data Access Working Group, 2008). Aware of the dichotomy of skills between music scholars and the ability to formulate such queries, the authors acknowledge that at present, the site UI presents notable barriers to access.

To address this, we have provided extensive documentation of the underlying ontological structures on the project website. Each of the sub-projects is illustrated with diagrams that include an interactive feature which provides the scope notes for a given class when hovered over with the cursor (see Figures 1-3). The diagrams also show the inherent connectivity of the graphs within *JazzCats*, and the directionality of properties (arrows running from domain to range). The combination of easy access to appropriate documentation for the data, the underlying ontological structure, and examples of functioning queries enables specialists to access *JazzCats* data according to their research interests. In addition, a Pubby interface (Cyganiak and Bizer, 2011) serves as an alternative Linked Data front end that enables users to navigate through the triples without the need to use SPARQL. Influenced by ongoing work at ResearchSpace (Oldman et al., 2016), planned future developments of *JazzCats* include the development of a GUI that will allow users to generate ontologically valid queries using dropdown lists generated by available properties for each class. This step will further help open *JazzCats* for experts and scholars along the full length of the digital humanities spectrum.



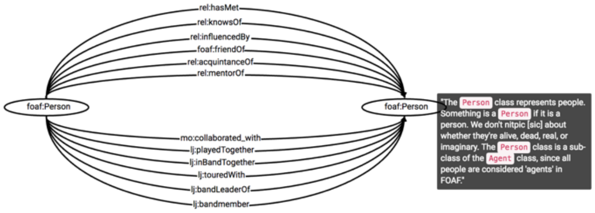
Subclass of MusicalExpression, which is defined as "The intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation, sound, etc., or any combination of such forms".

Figure 1. Body&Soul data structure illustrating pop-up scope notes for the class mo:Sound



Provides bibliographical information about a composition, such as title, artist, label.

Figure 2. WJazzD data structure illustrating pop-up scope notes for the class jcv:composition_info



"The Person class represents people. Something is a Person if it is a person. We don't nitpic [sic] about whether they're alive, dead, real, or imaginary. The Person class is a sub-class of the Agent class, since all people are considered 'agents' in FOAF."

Figure 3. LinkedJazz data structure illustrating pop-up scope notes for the class foaf:Person

Bibliography

Bangert, D. (2016). JazzCats Body and Soul discography. Zenodo. Dataset. http://doi.org/10.5281/zenodo.163886

Bangert, D., Nurmikko-Fuller, T. and Abdul-Rahman, A. (2016). *JazzCats.* Available at https://jazzcats.oerc.ox.ac.uk. Date last accessed: 31 Oct 2016.

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., ... & Gamble, M. (2013). "Why linked data is not enough for scientists." *Future Generation Computer Systems*, 29(2): 599-611.

Berners-Lee, T. (2006). Linked Data. Available at https://www.w3.org/DesignIssues/LinkedData.html. Date last accessed: 31 Oct 2016.

Bowen, J. (2013). Body and Soul discography. Available at http://josebowen.com/body-and-soul/. Date last accessed: 31 Oct 2016.

Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification 0.99. Namespace Document 14 January 2014. Paddington Edition.

Cyganiak, R. and Bizer, C. (2011). Pubby: A Linked Data Frontend for SPARQL Endpoints. Available at http://www4.wiwiss.fu-berlin.de/pubby/. Date last accessed: 6 April 2017.

Cyganiak, R. and Bizer, C. (2012). D2RQ: Accessing Relational Databases as Virtual RDF Graphs. Available at http://d2rq.org/d2r-server. Date last accessed: 31 Oct 2016.

Janowicz, K., Hitzler, P., Adams, B., Kolas, D. and Vardeman, I. (2014). "Five stars of linked data vocabulary use." *Semantic Web,* 5(3): 173-76.

Miles, A., Matthews, B., Wilson, M. and Brickley, D. (2005). "SKOS Core: Simple Knowledge Organisation for the Web."

Proceedings of the International Conference on Dublin Core and Metadata Applications, pp. 3-10.

Miller, P., Styles, R. and Heath, T. (2008). "Open Data Commons, a License for Open Data." *Linked Data on the Web*, 369.

Nurmikko-Fuller, T. (2016). SPARQL_queries_JazzCats. Zenodo. Dataset. http://doi.org/10.5281/zenodo.163879

Nurmikko-Fuller, T., Dix, A., Weigl, D.M. and Page, K.R. (2016). "In Collaboration with In Concert: Reflecting a Digital Library as Linked Data for Performance Ephemera." Proceedings of the 3rd International workshop on Digital Libraries for Musicology, pp. 17-24.

Oldman, D., Anagnostopoulou, M., Eales, G., Kelly, M. and Rychlik, A. (2016). ResearchSpace. British Museum. Available at http://www.researchspace.org. Date last accessed: 31 Oct 2016.

OpenLink Software, (2016). Virtuoso Universal Server. Available at https://virtuoso.openlinksw.com/. Date last accessed: 31 Oct 2016.

OpenRefine, (2013). OpenRefine 2.6 beta 1. Available at http://openrefine.org/download.html. Date last accessed: 31 Oct 2016.

Pattuelli, C. (2016). LinkedJazz Project. Available at https://linkedjazz.org/. Date last accessed: 31 Oct 2016.

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W-G., Burkhart, B. and Bartel, F. (2016). The Jazzomat Research Project, Doc v1.0. Available at http://jazzomat.hfm-weimar.de/dbformat/dboverview.html. Date last accessed: 31 Oct 2016.

Raimond, Y. and Abdallah, S. (2007). The Event Ontology. Technical report, Citeseer.

Raimond, Y., Abdallah, S. A., Sandler, M. B. and Giasson, F. (2007). "The Music Ontology." *Proceedings of the 8th International Conference on Music Information Retrieva*l, pp. 417-422.

RDF Core Working Group, (2014). RDF: Resource Description Framework. Available at https://www.w3.org/RDF/. Date last accessed: 31 Oct 2016.

University of Southern California, (2016). Karma: A Data Integraiton Tool. Available at http://usc-isi-i2.github.io/karma/. Date last accessed: 7 April 2017.

W3C RDF Data Access Working Group, (2008). SPARQL Query Language for RDF. Available at https://www.w3.org/TR/rdf-sparql-query/. Date last accessed: 31 Oct 2016.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... and Bouwman, J. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data,* 3.

Throughlines: Exposing Activism and Social Justice Issues in Los Angeles Transportation History

Britt S. Paris
parisb@ucla.edu
UC Los Angeles, United States of America

Marika Cifor
mcifor@ucla.edu
UC Los Angeles, United States of America

Protesters' shouts of "no justice, no peace" ring out over the rows upon rows cars forced a standstill on the 405, one of Los Angeles' busiest highways, on a July night in 2016. This recent protest by Black Lives Matter is notable and representative of an emergent social movement strategy. More than more than half of the 1400 protests relating to Black Lives Matter movement in nearly 300 U.S. and international cities from August 2015 to November 2015 effectively shut down transportation infrastructure (Badger, 2016). This contemporary activist practice can be seen as a logical tactic that shares roots with the historical occupations of schools, restaurants and administrative offices that occurred during protests in the civil rights movement in the United States in the 1960s and 1970s, as an earlier generation of activists rallied against racial segregation. Our proposed project makes accessible for the first time the history of transportation in Los Angeles and demonstrates through interactive maps, archival documents and audiovisual materials, as well as recorded oral history interviews why Los Angeles highways have become a productive site of protest. Contemporary highways are historically situated sites of contestation in which previous generations of racialized communities have paid the high price for Los Angeles' development into a renowned center of both commerce and culture.

Historian Eric Avila's (2014) *Folklore of the Freeway: Race and Revolt in the Modernist City* focuses on the racist highway projects that targeted and sought to isolate minority communities. With origins in the Jim Crow South, he argues, "federally funded highways were instruments of white supremacy, wiping out black neighborhoods with clear but tacit intent" (p. 43). Post-War Los Angeles, like many American cities, saw a rise in the prosperity of white, middle-class Americans who racialized desires and fears led them to towards suburbia in Los Angeles county and beyond. Redlining, restrictive covenants, and other practices of racial discrimination kept the dream of suburban home ownership out of reach of racial minorities. The new American dream came hand in hand with another new phenomenon, traffic. Highways, according to urban planning historian Joseph DiMento (2009), were recommended by urban planners as "the greatest single element in the cure of city ills. The federal government had stepped in by 1956 to cover the construction costs for highways by up to 90 percent. Much of this federal funding was also used by cities to rebuild and "redeem" urban areas (Semuels, 2016). New highways were not only placed not only as to move residents easily into suburbs and to other cities in the region, but were strategically planned and right placed by cities to eradicate what were termed "slums" and "blight," areas heavily populated by persons of color and the poor

(Semuels). Thus highways came to break apart rich and long-standing communities of color in Los Angeles and throughout the United States.

From the Chicano artists' depictions of the highway in East Los Angeles to second-line jazz parade in New Orleans, Avila illustrates how performative activities pleasure and protest by inhabitants remediate the spaces near highways to promote and reflect their own diverse perspectives, practices, and lived realities. East Los Angeles serves an important example of a racialized working-class community threatened by the construction of freeways. In the 1960s, the 60, 10, 101, 710, and 5 Freeways were all extended to cut through the East Los Angeles neighborhood. This environment spawned what Raul Homero Villa has coined as East LA's "expressway generation," from where some of East LA's finest muralists emerged. Throughout the 1970s, ASCO, one of the area's most important Chicano art collectives, used the walls of the freeways as a canvas to paint political slogans, like "Pinchi Placa Come Caca" (Fucking Pigs Eat Shit), "Gringo Laws = Dead Chicanos", "Kill the Pigs", and "Comida Para Todos" (Food For Everyone). ASCO turned these geographical sites of state power, the freeways, into forms of communication that expressed the relationship between spatial formation and racial tension.

The Black Lives Matter movement in Los Angeles has strategically returned to freeways. The press release for a December 23, 2015 protest that shut down a significant section of the 405 Freeway in the Westchester neighborhood explains the reasoning behind blocking freeway traffic. It states, "On one of the busiest travel days of the year, Black Lives Matter is calling for a halt on Christmas as usual in memorial of all of the loved ones we have lost and continue to lose this year to law enforcement violence without justice or recourse" (McReynolds, 2015). Pete White, an organizer with Black Lives Matter and the L.A. Community Action Network, told local news, "In this Christmas season, we're saying there is no mistletoe in our neighborhood, and it's not going to be business as usual" (McReynolds, 2015). Black Lives Matter actions in 2013, 2014, 2015 and 2016 have all resulted in successful and strategic blockages to major thoroughfares, clogging the arteries of the city, and bringing major media and popular attention. The tactical retaking of freeways in a notoriously automobile-driven city is symbolically and materially significant.

We will develop an online interactive repository to provide new access to a descriptive and underrecognized history of political contestation including the urban renewal movements of the mid-20th century, the activism of the expressway generation, and today's Black Lives Matter protests. The necessity of creating and animating this history is most clearly highlighted by the fact that there exists no consolidated archive of the historical images, dates and events related to this phenomenon. Our goal with this project is to provide a resource that would first, document the intersections of freeways, racial justice, and urban social movements in an easily-accessible website and second, promote sustained research and activism relating to this topic. Users are able curate their own path through the repository, for example, by navigating through the map, or scrolling through archival records linked to particular subjects arising within an oral history. This history will be illuminated through the activation of archival documents, video footage, and oral history interviews with individuals who have been involved in the historical development and issues of access intertwined with Los Angeles highways.

Bibliography

Avila, E. (2014) Folklore of the Freeway: Race and Revolt in the Modernist City. University of Minnesota: Minneapolis.

Badger, E. (2016) "Why Highways Have Become The Center of Civil Rights Protest" *The Washington Post.* (13 July, 2016). https://www.washingtonpost.com/news/wonk/wp/2016/07/13/why-highways-have- become-the-center-of-civil-rights-protest/

Dimento,J. F. (2009) "Stent (or Dagger?) in the Heart of Town: Urban Freeways in Syracuse, 1944-1967," Journal of Planning History, Vol. 8, No. 2: 133-161.

McReynolds, D. (2015) "Section of 405 Freeway Shut Down by Black Lives Matter Activists." *LAist.com* (December 23, 2015). http://laist.com/2015/12/23/section_of_405_freeway_shut_down_by.php

Semuels, A. (2016)"The Role of Highways in American Poverty," *The Atlantic (*March 18). http://www.theatlantic.com/business/archive/2016/03/role-of-highways-in-american-poverty/474282/

The Psalter Project: Providing Mediated Access to Religio-Political Subjects in Early Modern England

Nandra Perry
nandraperry@tamu.edu
Texas A&M University, United States of America

Bryan Paul Tarpley
bptarpley@tamu.edu
Texas A&M University, United States of America

Emerging from the crucible of the religious upheaval that characterizes the English Renaissance is arguably the most influential English book ever printed in terms of its impact on Anglophone religious, literary, popular, and legal culture: *The Book of Common Prayer* (BCP). Encoded within its pages is a kind of algorithm, an annually recurring process, a *ritualization* of both private devotion and public worship for generations of post-Reformation English readers. Taking our cue from Brad Pasanek (*Metaphors of Mind:*

An Eighteenth-Century Dictionary, 2015) and Peter Stallybrass ("Against Thinking." PMLA, Vol. 122, No. 5), both of whom have drawn useful analogies between the database and the commonplace book, we employ the creative anachronism of the "Bible app" to describe the function of the BCP in early modern England. As the first such "app" of its kind, the BCP choreographed religious meaning and ritualized worship for a whole generation of English Bible readers, shaping them into religio-political subjects who were then able to situate their lived experiences within a communally shared time and space. From the perspective of the Early Modern layperson, the BCP provides mediated access to the newly translated biblical text. Of course, from the abstracted perspective of the nascent nation-state of England, the BCP functions as a way to mitigate new anxieties surrounding the democratization of sacred scripture. As the legally established, official means by which sacred text is encountered, the BCP is nothing less than a masterpiece of social engineering.

To extend the metaphor of text as program, the BCP can also be thought of as a class, one which can be inherited and sub-classed, instantiated and "hacked" according to the agenda of particular readers who would produce, via their nuanced reading of BCP ritual, slightly different kinds of subjects according to the specific context in which they find themselves. Of particular interest to the project at hand is a 1586 BCP which has been highly "sub-classed" by one of its owners. Bound together with the prayer book is an entire psalter, whose collection of 150 psalms is cross-referenced in a single hand, which also makes occasional thematic/tonal annotations. In our examination of this prayer book, we wish to develop a methodology for accessing the kind of subject such a "re-engineered" BCP might have produced.

Implied in the very notion of access, of course, is mediation. Within the limited scope of our project, we do not have recourse to the intense amount of labor required to perform a rigorous exegesis of the entire psalter according to how its 16th-century readers might have read it. What we do have, via the psalter's marginalia, is what one (or perhaps two) reader(s) selected as noteworthy in their BCP-regulated practice of reading the Psalms. We also have our own attempts to thematize and register the tone of those same texts. Given these assets, we attempt to provide via the Psalter Project a representation of how a subject produced by this prayer book might look *from our perspective*. Our hope is that, despite the inherently mediated nature of such a representation, we might provide students and scholars alike a better understanding of the "programmatic" nature of religious para-texts like the BCP.

The Psalter Project was born out of Dr. Nandra Perry's scholarship in Early Modern English literature in partnership with Bryan Tarpley's work as Lead Software Applications Developer for the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University. Perry's work with the 1586 prayer book led her to apply for a Summer Technical Assistance Grant with the IDHMC, at which point she was awarded some of Tarpley's development time. Together, they (we) designed a relational database schema for recording the cross-references and tags found in the 1586 prayer book, as well as a web-interface for entering and viewing them. A beta version of the project is available for viewing.

Before commencing development on the Psalter Project web app, a survey of extant, web-based tools was performed to determine whether any one tool (or collection of tools) already satisfied our requirements. Given the marginalia we have to work with (a large amount of scriptural cross-references and thematic tags), we wanted a tool to facilitate the capture and analysis of this data. While there are indeed several digital annotation tools available, such as MIT's Annotation Studio, or the University of Virginia's PRISM, none of these tools allow for the rapid entry of scriptural verse range references and tags. Tarpley implemented the database schema by creating a MySQL database and wrote the web application in Python using the Django web framework, as well as other web-related technologies, such as the jQuery Javascript library and the Bootstrap CSS framework. The database and web app are both hosted by the IDHMC's server infrastructure at Texas A&M. As a way of beta-testing the app, Perry then recorded a sampling of the psalter's cross-references and tags, and also included her own tagging of the referenced verses in terms of both thematic content and perceived emotional impact (affect).

While the Psalter Project in still in development both in terms of methodology and finished product, a preliminary sketch of our reader is emerging via the various views made available through the web app. There is a visualization of the thematic and affective tags in the form of tag-clouds weighted by frequency, a view of the most frequently referenced verses along with their tags, a break-down of all referenced verses (and associated tags) by any specific thematic or affective tag, a presentation of the network of referenced verses (and associated tags) by any specific verse, and a presentation of any of the 150 Psalms in its entirety where referenced verses can be hovered over, displaying tags and referenced verses. At this initial stage, we believe that this multi-faceted portrayal is revealing a reader with a profound sense of group identity as the loyal subject of a just God who provides deliverance to the deserving and punishment to "bad" subjects and oppressive outsiders. The effect, it would seem, is the sacralization of a religio-politics in which the reader's relationship to God is analogous to his/her relationship to the nation-state (and vice versa), thereby justifying, in turn, a posture of hostility toward outsiders and "bad" subjects.

With this paper, we intend to not only provide a more fully fleshed-out representation of the Early Modern religio-political English subject, but to interrogate the various assumptions and methodologies we use to provide this representation so that we might improve the Psalter Project web application. We hope to be able to provide this web application (both in terms of an open- source repository and as an IDHMC hosted web service) to other scholars in the future, so that they too might be able to provide (mediated) access to religio-political subjects.

Medieval Textual Transmission Modeling in Unity3D

Lynn Ramey
lynn.ramey@vanderbilt.edu
Vanderbilt University, United States

How did texts and ideas circulate within and between societies in the Middle Ages? We know that there were many potential vectors of movement: pilgrimage, crusade, merchant caravans and ships, and itinerant performers, to name a few. However, particularly in societies where transmission was largely oral, scholars usually cannot identify specific moments and locations when stories moved from one location to the next. For the medieval period, this lack of data has historically been the cause of heated debates as scholars identify stories that share common elements, but due to historical or political reasons they are resistant to any notion that one nation's literary tradition is "indebted" to another. Nonetheless, societies are influenced by art and ideas from around the world, and denying the contributions of global societies to stories that are held dear leads to isolationism and a sense of cultural superiority. In 2016, US Congressman Steve King (R-Iowa) questioned what non-whites have contributed to civilization. I am presenting an early version of a project that aims to address a part of that question, giving an immersive digital experience of what some of those contributions might look like and just when and how they occurred.

My project uses the Unity game engine to create stories of textual transmission. In particular, I am modeling how elements of the Thousand and One Nights could have circulated between East and West via the Lusignan court at the crusader kingdom of Cyprus circa 1194. In collaboration with Professor Sahar Amer (Arabic Studies, University of Sydney), we have identified a story from the Arabic text, The Prince Qamar Al Zaman and Princess Boudour, that shares narrative patterns and tropes with Old French romances *Floire et Blanchefleur, L'Escoufle, Huon de Bordeaux*, and *Miracles de la fille d'un roi.* Of course, each story has unique elements, but some of the striking scenes (dramatic public unveiling of mistaken gender identity) and themes (a princess who cross dresses and becomes an itinerant knight search of her lover only to be such a valiant warrior that she winds up married to her lord's daughter) are repeated. What is the relationship between Princess Boudour and Blanchefleur? Circumventing the theories of Joseph Campbell and Vladimir Propp that stories have basic, shared, "universal" components, we look at cultural context and emphasize variants. Our digital story will research and test various proposed vectors of transmission that ultimately result in the morphing of one story into a different incarnation of the same tale. While we will never have all versions of a story with changes and variations as it moved from one culture to the next, we can find the smoking gun as it were— places, times, and persons that we know participated in the exchange of objects, texts, and people. By making these moments visible we can test theories and vectors of transmission.

For example, in the medieval period stories were often changed slightly to please the patrons who paid for them. Traveling minstrels would insert the names of local lords or knights into the story, likely to increase eventual payment at the end of the evening. Stories might also change to use more familiar geographical locations. More dramatic changes could find their way into via mistakes in translation or understanding of stories that passed from one court to the next. Some scribes and minstrels enjoyed changing tales drastically in order to leave their own mark. Within our 3D environment, users will experience the multicultural and multilingual Lusignan court and its environs. Users must make small decisions that will ultimately result in changes, sometimes dramatic, in the final tale.

By illustrating just a few of the factors that influence storytelling and cultural interaction, we conceptualize two audiences. The first is someone who may not have ever considered how much cultural interchange can happen in simple storytelling. As the user makes choices, he or she sees and experiences directly the storytelling environment, becoming agents in the creation process. Our second audience is the researcher who, while being abstractly aware of all of these factors, may not see the potential impact of textual manipulation on a story important to his or her research. By working through these connections in an immersive environment, researchers will consider the global nature of the movement of ideas and interchanges in other texts.

The project brings together two areas of academic interest: global medieval studies and spatial studies. Both areas have well-developed audiences.

1. This project is part of the Global Middle Ages Project that aims to make apparent the global cultural interactions that took place in the medieval period, a time that is frequently studied in nationalistic isolation with an apparent assumption that before the modern period people did not travel and were not aware of the wider world around them. With its strategic location in the Mediterranean, Cyprus served as a crossroads between east and west in the Middle Ages. Despite their disparate backgrounds, the Cypriots lived together on this relatively small island for centuries. Scholars (Andrews, Carr, Grünbart, Nicolaou-Konnari, Rogge, Schabel, and others) have looked at this multi-cultural space, establishing moments of contact and exchange using artifacts of architecture and art history. We will extend their work using data gathered from literary works.

2. As we create a space where users can experience the past, we must consider that the meaning and experience of that space was likely very different 1000 years ago. Robert Tally, following and modifying the work of Bertrand Westphal, Leonard Goldstein, and others, suggests that perceptions of time and space changed radically about the time that linear perspective arose in artistic works. In a medieval illumination, for instance, one might have snapshots of past, present, and future within one framed block. Foreground and background rest in the same plane. Starting with perspectival drawing, Goldstein suggests that space became experienced as 1) continuous, isotropic, and homogenous; 2) quantifiable; and (3) perceived from the point of view of a single, central observer. As a contrast, a medieval map such as the 1375 Catalan Atlas evinces a different notion of time and space, at least in terms of visual representation. The map simultaneously includes the Queen of Sheba, the three wise men, and the Great Khan from vastly disparate eras (Old Testament, New Testament, and thirteenth century). Distances and landmasses are not measured to a modern sense of scale, though meaning can be induced from the relative sizes given to different areas and cities. Furthermore, we cannot know from which end the viewer was meant to see the map, as the writing and images are oriented in all directions. Space could comprise multiple times and places, and the lack of orientation toward a particular viewer indicates that perhaps multiple perspectives were expected.

In this sense of space, medieval maps share many traits of video game maps, including the relationship between place and object, where time, distance, and proportion vacillate between being important and irrelevant. For Mario in Super Mario Bros., the distance on the map is not proportionate to the difficulty of the task of reaching the next point. On the Catalan Atlas, Sheba and the Three Wise Men are a part of the map because they share space and time since Creation, showing that time unites us (we are all postlapsarian and part of the map of Christian history) and separates us (we can never meet up since we do not exist at the same moment). A video game engine provides a superlative space for modeling these pre-modern notions of time and space that tell us much about how people, ideas, and texts circulated in the medieval Mediterranean.

Bibliography

Andrews, J. (2012)"Conveyance and Convergence: Visual Culture in Medieval Cyprus." *Medieval Encounters* 18: 413–446.

Champion, E. (2011*). Playing with the Past.* London: Springer-Verlag

Ellis, R. (1994) "Textual Transmission and Translation in the Middle Ages." Ed. Elizabeth Archibald et al. *Translation and Literature* 3 (1994): 121–130.

Ellwood, R. (1999) *The Politics of Myth: A Study of C. G. Jung, Mircea Eliade, and Joseph Campbell*. Albany: SUNY Press, 1999.

Lefebvre, H.(1991) *The Production of Space.* Trans. D. Nicholson-Smith. Oxford: Blackwell, 1991.

Ramey, L., and Panter, R. (2015). "Collaborative Storytelling in Unity3D: Creating Scalable Long-Term Projects for Humanists." *Interactive Storytelling*. Ed. H. Schoenau-Fog et al. Vol. 9445. Springer-Verlag, 357–60.

Rogge, S., and Grünbart, M. (2015). *Medieval Cyprus: A Place of Cultural Encounter.* Waxmann Verlag,

Rowland, T. (2014) "We Will Travel by Map: Maps as Narrative Spaces in Video Games and Medieval Texts." *Digital Gaming Re-Imagines the Middle Ages*. Routledge. New York, NY: Daniel T. Kline.189–201.

Tally, R. T. (2013) *Spatiality*. New York: Routledge, 2013.

A Shared Task for a Shared Goal: Systematic Annotation of Literary Texts

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
Stuttgart University, Germany

Evelyn Gius
evelyn.gius@uni-hamburg.de
Hamburg University, Germany;

Jannik Strötgen
jannik.stroetgen@mpi-inf.mpg.de
Max Planck Institute for Informatics, Germany

Marcus Willand
marcus.willand@ilw.uni-stuttgart.de
Stuttgart University, Germany

Introduction

In this talk, we would like to outline a proposal for a shared task (ST) in and for the digital humanities. In general, shared tasks are highly productive frameworks for bringing together different researchers/research groups and, if done in a sensible way, foster interdisciplinary collaboration. They have a tradition in natural language processing (NLP) where organizers define research tasks and settings. In order to cope for the specialties of DH research, we propose a ST that works in two phases, with two distinct target audiences and possible participants.

Generally, this setup allows both "sides" of the DH community to bring in what they do best: Humanities scholars focus on conceptual issues, their description and definition. Computer science researchers focus on technical issues and work towards automatisation (cf. Kuhn & Reiter, 2015). The ideal scenario– that both "sides" of DH contribute to the work in both areas– is challenging to achieve in practice. The shared-task scenario takes this into account and encourages Humanities scholars without access to programming "resources" to contribute to the conceptual phase (Phase 1), while software engineers without interest in literature per se can contribute to the automatisation phase (Phase 2). We believe that this setup can actually lower the entry bar for DH research. Decoupling, however, does not imply strict, un-crossable boundaries: There needs to be interaction between the two phases, which is supported by our mixed organisation team. In particular, this setup allows mixed teams to participate in both phases (and it will be interesting to see how they fare).

In Phase 1 of a shared task, participants with a strong understanding of a specific literary phenomenon (literary studies scholars) work on the creation of annotation guidelines. This allows them to bring in their expertise without worrying about the feasibility of automatisation endeavours or struggling with technical issues. We will compare the different annotation guidelines both qualitatively: by having an in-depth discussion during a workshop, and quantitatively: by measuring inter-annotator agreement. This will result in a community guided selection of annotation guidelines for a set of phenomena. The involvement of the research community in this process guarantees that heterogeneous points of view are taken into account.

The guidelines will then enter Phase 2 to actually make annotations on a semi-large scale. These annotations then enter a "classical" shared task as it is established in the NLP community: Various teams competitively contribute systems whose performances will be evaluated in a quantitative manner.

Given the complexity of many phenomena in literature, we expect the automatisation of such annotations to be an interesting challenge from an engineering perspective. On the other hand, it is an excellent opportunity to initiate the development of tools tailored to the detection of specific phenomena that are relevant for computational literary studies.

This talk has two purposes:

- To discuss these ideas and collect feedback and propositions. This is also an explicit invitation to contribute in the setup of this initiative. We are also welcoming a discussion about the phenomena that should be included.
- To advertise the idea of a shared task and to invite possible participants. The success of STs relies on a certain number of participants. Given that this has never been organized in the DH community before, we want to spread this idea throughout the community to gather estimates of potential participants.

The Importance of Annotations

In computational literary studies, many phenomena cannot directly be detected from the text surface. To find and categorize such phenomena as, for example, the "narrated time" in a novel, it is first necessary to have an in-depth understanding of the text, knowledge about its author or literary conventions, or knowledge of the text's historical context. Therefore, instances of such phenomena need to be annotated either by human experts or software that is tailored to this task.

Unfortunately, many theories describing interesting phenomena are very difficult to apply to real texts. It has been shown numerous times (e.g., Reiter, 2015, Musi et al., 2016) that annotating theories or concepts directly can lead to very poor inter-annotator agreement (IAA): Different annotators have different interpretions of not only the text, but also descriptions of the theoretical concepts. Although subjective annotations have their merit, studying annotations on large scale depends on their consistency, i.e., a high IAA. In addition, many theories are underspecified and provide examples for illustrations only. Creators of annotation guidelines often have to interpret what is meant by a certain statement and extend definitions to cover examples found in real texts.

Annotation guidelines serve as a mediator between the annotators and a theory (that may use specialised vocabulary). Additionally, such guidelines often contain re-appearing instance patterns and their modes of annotation and/or exceptions, as well as many examples from real texts (see below).

We see the creation of annotation guidelines as one of the cornerstones of large scale text analysis in computational literary studies. Additionally, the creation of annotiation guidelines supports systematic disciplinary discussions about concepts and thus may lead to additional findings relevant for the theoretical discourse (e.g., Meister, 1995; Gius and Jacke, forthcoming). Experts from the field literary studies are well-suited to work with annotation guidelines, as annotation of literary phenomena in literary texts can be seen as a special form of close reading.

Phase One: Annotation Guidelines

The Shared Task

In theory, any phenomenon can be addressed in this fashion, as long as it can be defined inter-subjectively, is reasonably frequent, and is of interest in computational literary studies. As a starting point, we propose to address the issue of narrative levels (Pier, 2014). Narrative levels are a core concept in narrative theory (Genette, 1980; Bal, 1997) which in turn has shown to be a promising foundation for automatisation in literary theory (Bögel et al. 2015). The first reason for choosing to examine narrative levels is their ubiquity: every narrative text necessarily consists at least one, most texts contain many narrative levels; each element of a text can be assigned to a specific level. The second reason is our intuition that a definition of "level" in guidelines is as achievable as the automated detection of levels by computers.

Concretely, participating teams are asked to create guidelines for the detection and annotation of a) narrative levels and b) the relation of the narrator(s) to the narrated world (i.e., is the narrator part of the narrated world or not?). Participants are not bound to adhere to a specific narratological theory. The result of this phase, however, will be a fixation on a set of guidelines (that instantiate a theory).

We will select a number of literary narrative texts and provide copyright-free digitized corpora. All “official” texts (development and test sets) will be English literary texts. Naturally, the second step will be to extend this framework to other languages and/or phenomena.

Evaluation

In NLP shared tasks, the predictions of the systems are compared against a fixed test set– the “gold standard”. Since there is no gold standard in Phase 1, we will evaluate the guidelines using an unseen data set. Each participating team annotates this set using their own guidelines before the guidelines are submitted. Submitted guidelines will be anonymized and re-distributed among the participants. Each participant is asked to annotate the evaluation data set using two other annotation guidelines. In addition, we will be collecting annotations from students.

The evaluation data set will thus be annotated according to each participant’s guidelines four times (1x self, 1x student, 2x other participants).

This setup allows direct calculation of inter-annotator agreement. However, IAA should be only one aspect in evaluating the guidelines, but not the only one. Therefore, we will submit a workshop at DH 2018 to discuss the submissions and select final ones. This setup also allows the merging of different annotation guidelines as well as adaptation according to the discussion during the workshop. Soon after the workshop, Phase 2 will commence.

Phase Two: Automatic Prediction

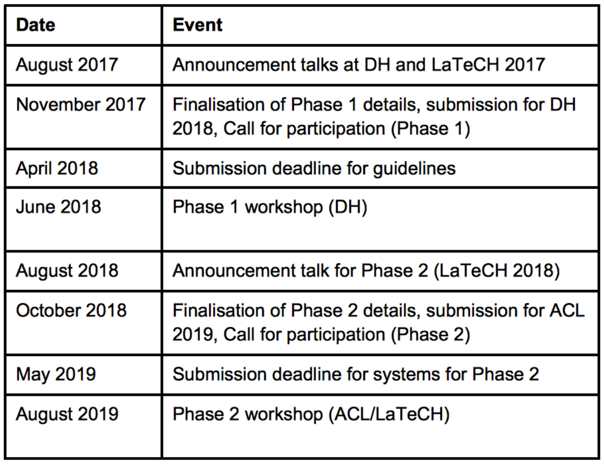
In Phase 2 of this endeavour, the selected guidelines of Phase 1 will be annotated on a large scale by student assistants. Since we do not yet know how long, complex and involved the guidelines will be, there should be close communication between the organizers and the team responsible for the selected guidelines.

As soon as the texts have been annotated, they will be made accessible to participants. For the deadline in Phase 2, participants will be asked to process a new data set – one that has not been released before – and return the predicted annotations to the organizers, who will then make evaluations using standard measures (e.g., accuracy). Finally, there will be a workshop in which each participating team presents its system. This workshop is projected to be coordinated with the LaTeCH workshop series, which has taken place at ACL conferences in the past years (one of the authors of this paper has been a workshop organiser in past years).

There will be no fixation on computational approaches, statistical models, programming languages or environments to tackle this problem. The main benefit of shared tasks towards automatisation is that different approaches can be compared directly. Restricting this possibility space would directly harm the goal.

Technical Details and Timeline

This proposal is innovative in a number of ways: Shared tasks are a new kind of framework in the DH/H community, as such, focalisation have not been investigated in this way before. Last but not least, a shared task with the goal of creating annotation guidelines has not been organized before (to our knowledge). We believe it is more important to do this right than to do this fast, hence we are looking at a rather lengthy timeline.



Date	Event
August 2017	Announcement talks at DH and LaTeCH 2017
November 2017	Finalisation of Phase 1 details, submission for DH 2018, Call for participation (Phase 1)
April 2018	Submission deadline for guidelines
June 2018	Phase 1 workshop (DH)
August 2018	Announcement talk for Phase 2 (LaTeCH 2018)
October 2018	Finalisation of Phase 2 details, submission for ACL 2019, Call for participation (Phase 2)
May 2019	Submission deadline for systems for Phase 2
August 2019	Phase 2 workshop (ACL/LaTeCH)

Attracting enough participants is the main challenge from the organiser's perspective. The main incentives we envision for contributors are excellent publication opportunities: All submitted and generated materials will be published online (open access) under the umbrella of the shared task. Each individual work will be citable. This includes the submitted annotation guidelines, the produced consensus guidelines, and explanation and commentary documents.

In addition to the publication incentive, we believe that our approach is an important contribution towards systematic text analysis in the DH realm. We count on the playfulness and curiosity (in the best sense!) of the DH community to take part in this experiment.

Bibliography

Bal, M. (1997). Narratology: Introduction to the Theory of Narrative. University of Toronto Press, 2nd edition.

Bögel, T, Gertz M., Gius, E., Jacke, J., Meister, J. C., Petris, M., and Strötgen, J. (2015). Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative. In DHCommons Journal, 2015. URL http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heurecl%C3%A9-digital-heuristic.

Carlson, L., and Marcu, D (2001). Discourse tagging reference manual. Annotation Manual, University of Southern California,

2001. URL https://www.isi.edu/ marcu/discourse/tagging-ref-manual.pdf.

Ferro, L., Gerber, L, Mani, I., , Sundheim, B, and Wilson, G. (2005). TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, The MITRE Corporation.

Genette, G. (1980). Narrative Discourse. An Essay in Method. Ithaca, N.Y: Cornell University Press.

Gius, E., and Janina Jacke, J. (2016). Zur Annotation narratologischer Kategorien der zeit. Annotation Manual 2.0, Hamburg University, November 2016. URL http://heureclea.de/guidelines

Gius, E., and Jacke, J. (n.d.)The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis. International Journal of Humanities and Arts Computing, forthcoming.

Hovy, E., and Lavid. J. (2010) Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. International Journal of Translation Studies, 22(1).

Kuhn, J., and Reiter, N. (2015) A Plea for a Method-Driven Agenda in the Digital Humanities. In Proceedings of Digital Humanities 2015, Sydney, Australia, June 2015.

Meister, J. C. (1995). Consensus ex Machina? Consensus qua Machina! Literary and Linguistic Computing, 10(4), pages 263–270.

Musi, E., Ghosh, D., and Muresan, S. (2016) Towards feasible guidelines for the annotation of argument schemes. In Proceedings of the Third Workshop on Argument Mining (ArgMining2016), pages 82–93, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-2810.

Pier, J. (2014). Narrative Levels (revised version; uploaded 23 April 2014). In Peter Hühn et al., editors, the living handbook of narratology. Hamburg: Hamburg University. URL http://www.lhn.uni-hamburg.de/article/narrative-levels-revised-version-uploaded-23-april-2014

Reiter, N. (2015). Towards annotating narrative segments. In Kalliopi Zervanou, Marieke van Erp, and Beatrice Alex, editors, Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 34–38, Beijing,China, July 2015. Association for Computational Linguistics, Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-3705.

Santorini, B. (1992) Penn treebank: Part-of-speech tagging. Annotation Manual 3, University of Pennsylvania, 1992. URL ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

Sauri, R., Littman, J., Knippen, B., Gaizauskas, R., Setzer, A., and Pustejovsky, J. (n.d.) TimeML Annotation Guidelines, Version 1.2.1. http://timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf

Styler, W., Savova, G., Palmer, M., Pustejovsky, J., O'Gorman, T., and de Groen, P. C. (2014) THYME Annotation Guidelines. Technical report. http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

Ngrams Against Agnotology: Combatting Tobacco Industry Narratives About Addiction Through A Quantitative Analysis Of 14 Million Documents

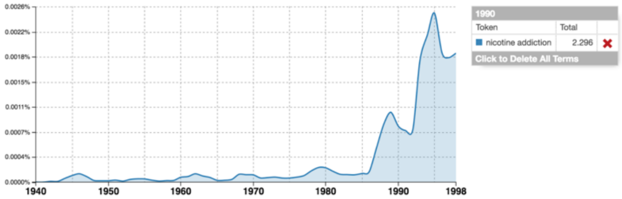
Stephan Risi
risi@stanford.edu
Stanford University, United States of America

What happens when researchers have access to more documents than they could read in a lifetime? As a result of litigation, historians of tobacco have access to over 14 million formerly secret tobacco industry documents, containing incriminating internal memos and research reports but also newspaper clippings and consumer letters (UCSF Library and Center for Knowledge Management). Historians have used this treasure trove to document widespread fraud and systematic deception of smokers by the tobacco industry

(Brandt, 2007; Proctor, 2011). However, industry-friendly historians and tobacco lawyers have started to rewrite some this history by claiming that smokers always knew that smoking is addictive and causes cancer (Brandt, 2007; Proctor, 2011). Usually, these claims rest on a few, well selected documents that support a particular industry claim. Robert Proctor has called processes like this "agnotology," the cultural production of ignorance (Proctor, 2008). Indeed, the arguments of both pro- and anti-tobacco industry historians rely on the same corpus of data: an immense amount of publicly available and full-text searchable documents. Given 14 million documents, there will be some supporting almost any claim.

In this paper, I present one way to counter such agnotological assertions by studying broad trends across millions of documents with frequency analyses. In particular, I counter the claim that smokers always knew that smoking was addictive, an argument often made by tobacco lawyers in court to assign full responsibility to the smoker (Henningfield, Rose, & Zeller, 2006) To refute this assertion, I use frequency analyses with a validation measure to show that smoking only became widely understood as an addiction in the late 1980s and early 1990s, when scientists recognized that the same neural pathways were involved in dependence to both nicotine and harder drugs like heroin and cocaine. This inscription of addiction into the brain replaced older explanations of why people smoke, like personality traits or an oral fixation. Ultimately, I trace how the neurological understanding of nicotine addiction moved from research laboratories to the public: it led to the Surgeon General's warning labels; it enabled smokers to seek out new nicotine replacement therapies; and it made it possible for smokers to successfully sue the tobacco industry for the first time.

Frequency analysis, popularized by the Google Ngram Viewer, is one of the simplest mathematical tools in the arsenal of the digital humanities (Michel et al, 2011). By calculating the usage frequency of terms and expressions over time, it enables users to get a sense of when a term became more or less important. The mathematical simplicity confers it an important advantage: it scales very well not just to thousands but to millions of documents. For this study, I used all 14 million documents (about 10 billion tokens) dated between 1940 and 1998 to create a publicly available website (Tobacco Analytics), where users can create their run their own frequency analyses, akin to Google ngrams.



Token	Total	
■ nicotine addiction	2.296	✖

Figure 1: Screenshot of the relative frequencies of "nicotine addiction" in the tobacco documents from www.tobacco-analytics.org. The presentation will use a number of these graphs to show that smoking only became understood as an addiction in the late 1980s and early '90s.

The main drawback of this method is that the patterns found in the graphs of the Google Ngram Viewer are hard to validate: Does a spike in a particular year represent a statistically significant event or is it just a fluke? Is it caused by 10 or 1000 documents? I address this problem in two ways: First, I allow users to display the absolute number of appearances of a term by year to give them a sense of the number of documents that cause a spike. Second, I am developing a comparison statistic to calculate z-scores using the Corpus of Historical American English (COHA) (Davies, 2010). By comparing frequencies between the tobacco documents and the reference corpus (COHA), it allows me to calculate when frequencies in the tobacco corpus deviate in a statistically significant way (Darwin, 2008, p. 208-222). Given, for example, the above graph of the relative frequencies of the term "nicotine addiction," z-scores can be used to show that the relative frequencies only started to deviate significantly from the comparison corpus in the 1980s.

The tobacco documents provide us with an opportunity to think through the problems that come with access to millions of secret documents. What if millions of dollars in settlements hinge on historical arguments? What if there are immense financial incentives to make false historical claims: to present narratives that are borne out in a few well selected documents, but which misrepresent the corpus as a whole? In the realm of tobacco, historical arguments and knowledge circulate far outside of academia in courtrooms to sway juries or in policy documents to change legislation. The immense size of the tobacco documents archive makes it possible to find a few documents supporting almost any claim. Findings from one group of documents can cancel out the findings from other documents; statements by one expert discredit those of another one. In these cases, quantitative analyses using the whole corpus can be an arbiter of these claims. They are not sufficient to advance historical arguments in themselves, but they can be used to test and disprove hypotheses made on the basis of a smaller set of documents. The Tobacco Analytics project makes powerful digital humanities tools available to tobacco researchers who may not have a technical background, and it allows historians to trace developments within the tobacco industry by examining the whole corpus with the click of a mouse.

Bibliography

Brandt, A. (2007). *The Cigarette Century: The Rise, Fall, and Deadly Persistence of the Product that Defined America*. New York: Basic Books.

Darwin, C. M. (2008). *Construction and Analysis of the University of Georgia Tobacco Documents Corpus.* PhD Dissertation, The University of Georgia, Athens, GA.

Davies, M. (2010). The Corpus of Historical American English: 400 million words, 1810-2009.

Henningfield, J., Rose, C., & Zeller, M. (2006). Tobacco Industry Litigation Position on Addiction: Continued Dependence on Past Views. *Tobacco Control, 15*(Suppl. 4), 27-36.

Kyriakoudes, L. M. (2006). Historians' Testimony on "Common Knowledge" of the Risks of Tobacco Use: A Review and Analysis of Experts Testifying on Behalf of Cigarette Manufacturers in Civil Litigation. *Tobacco Control, 15*(suppl 4), iv107-iv116.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . Orwant, J. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331*(6014), 176-182.

Proctor, R. N. (2008). Agnotology. A Missing Term to Describe the Cultural Production of Ignorance (and Its Study). In R. N. Proctor & L. Schiebinger (Eds.), *Agnotology. The Making and Unmaking of Knowledge*. Stanford, CA: Stanford University Press.

Proctor, R. N. (2011). *Golden Holocaust: Origins of the Cigarette Catastrophe and the Case for Abolition*. Berkeley: University of California Press.

Integrating Image Resources Into Virtual Research Environments For The Humanities – a Simple Image Presentation Interface (SIPI) based on IIIF

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
University of Basel, Switzerland

Andrea Bianco
andrea.bianco@unibas.ch
University of Basel, Switzerland

Peter Fornaro
peter.fornaro@unibas.ch
University of Basel, Switzerland

Introduction

Presenting images on the web has a long tradition. The <img>-tag was proposed February 1993 by Marc Andreessen, at this time employed by the National Center for Supercomputing Applications of the University of Illinois at Urbana-Champaign, a pioneering institution for the development of the World Wide Web. Marc Andreessen later founded Netscape, the browser that helped to make the web popular. Its most simple form, the image tag contained an src-attribute containing the URL of the image, such as **<img src="http://dhlab.unibas.ch/imgs/sample.jpg">.** Basically, the mechanism to include images into a website has since remained the same. The dynamic manipulation of the DOM with JavaScript (also a invention introduced by Marc Andreessen and Netscape) allowed users to dynamically add, delete or exchange images within a website. Thus complex web-applications that dealt dynamically with images became possible.

However, while on the side of the browser some scaling is possible, the URL had to address an image file with fixed properties (file format, image size etc.). As the computing power of servers grew, it became possible to render the image on the fly with properties that are passed using the HTTP request (e.g. embedded in the URL or as URL-parameters). Recently, the International Image Interoperability Framework (IIIF) emerged as a standard syntax to pass parameters such as image size in pixels, cropping region, rotation angle and file format using a well-defined URL syntax. The IIIF-standard allows to share image resources between web-applications such as Virtual Research Environments (VRE) in a standardized way. There are quite a few IIIF-compliant image servers around, some using a native implementation, some wrapping image transformation programs using a script language such as python.

By a mandate of the State Secretariat for Education, Research and Innovation (SERI), the Swiss Academy of Humanities and Social Sciences (SAHSS) has created a new institution for the preservation and long-term curation of research data in the humanities (Data and Service Center for the Humanities, DaSCH). It ensures permanent access to research data in order to make it available for further research. A pilot program started in 2013 and has been successfully finished. The DaSCH will be permanently installed on January 1st 2017, to guarantee seamless services. The Digital Humanities Lab (DHLab) of the University of Basel has been mandated with the operation of the new institution. For this purposed, DHLab has developed a flexible platform based on semantic web technologies (RDF, RDFS, OWL). Besides text sources, about 500,000 high-resolution images have been ingested to the system during the pilot phase. We decided to use IIIF for presenting the images in order to maximize the interoperability with external systems. Furthermore, we need to preserve only one image file, since IIIF allows using the archiving master also for dissemination and presentation. However, none of the existing IIIF-compliant servers satisfied our needs.

Design Requirements

Therefore, we decided to design and implement our own IIIF server guided by the following requirements:

- Interoperability with internal databases (e.g. RDF-triplestores, RDBMS etc.) containing annotations, metadata etc. as well as access permissions.
- Preservation of all embedded metadata (e.g. EXIF, XMP, TIFF etc.) during all format conversions
- ICC color profile conversions where necessary
- User authentication compatible with the proposed draft of the IIIF for authentication
- High-performance transformation of images including rotation, format conversions (e.g. 16 bit to 8 bit depth) etc.
- Support of Secure Socket Layer (SSL/https)
- A configurable image cache in order to reduce the computational load on the server
- Support of cross origin resource sharing (CORS)
- Import and transformation of images. The server must be able to import images and convert them to the desired master file format (in our case JPEG2000).
- Features beyond the scope of the IIIF-standard such as adding watermarks, size restrictions etc.
- Integrated simple webserver functionality
- Modular extensibility, e.g. integrating support for RTI imaging (both initial transformation and

serving a web-based RTI-viewer – Fornaro et al, 2016)

Implementation

In order to fulfill these requirements, we decided to use C++11 to develop a native, modular, high performance IIIF-compliant image server. A decisive feature was to make the server scriptable with a script language without compromising on performance. We decided to use Lua, an extremely fast, performing, and extensible script language with a small footprint that has only small overhead. It is widely in use for computer games such as World of Warcraft or Minecraft. In order to support different image file formats, we use open source libraries such as libtiff, libjpeg etc. and a modular, extensible architecture. In order to support the JPEG2000 image format, we rely on the kakadu-library. Unfortunately kakadu is not open source, but it is one of the most performant JPEG2000-libraries available. In addition, when acquiring a license, the full source code is provided. Depending on the licensing model, the free distribution of binary packages is included.

In order to reduce the computational load of the server, an efficient caching service has been implemented. The canonical IIIF-URL is used as key for caching since it is unique for each parameter set of the IIIF request.

The Lua-interpreter has been extended with SIPI-specific functionality. Using configurable routes, a fully IIIF-compliant image server has been implemented with the following features:

- Full support of SSL (https) using the OpenSSL library.
- Preservation of metadata. We use the open source exiv2 library and own parsing/generating routines to bridge the differences between the different image formats.
- ICC profiles are interpreted and converted using the open source "little cms" library.
- Before serving an image, a configurable pre-flight lua-script may be executed. Within this script different tasks can be performed, e.g. querying a database for access rights, adding watermarks, ICC profile conversions etc.
- Native support of JSON Web Tokens. JWT's may be analyzed using simple Lua functions (e.g. in the pre-flight script) for authentication and access control.
- Querying other databases using RESTful services. These RESTful query functions are also exposed to Lua.
- Image upload using HTTP multipart/form-data headers. The upload process and file conversions can be controlled with simple Lua scripts.
- Cache control with a simple web-based administration interface.
- Native support for sqlite3 databases from the embedded Lua.

SIPI is open source and can be found on Github. In order to use the JPEG2000 format, a licensed copy of the kakadu library has to be provided by the user. We will provide an extensive manual and binary downloads (including JPEG2000 support) for all major Linux distributions, OS X and a docker image on http://sipi.io

Conclusion

SIPI is a fully IIIF-compliant native image server which integrates extremely well into existing platforms. The flexibility provided by the embedded scripting language, as well as the features going beyond the IIIF specification, allow the integration of IIIIF-based interoperability into existing imaging platforms and image repositories. The support of the secure socket layer (https) access control is a necessity in the environment of digital humanities.

SIPI can easily be customized and extended for special purposes. Elaborated imaging methods (e.g. support for multi-spectral images, image processing functions etc.) could be implemented using the existing SIPI server as base. However, such enhancements will require extensions to the IIIF-syntax. Further development at the DHLab will include the preprocessing of RTI-images as well support for other media types such as PDF and moving image. We are also working on the implementation of the Amazon S3 client-API in order for SIPI to directly serve images that are stored in a S3 compatible cloud storage.

Bibliography

Fornaro, P., Bianco, A., Rosenthaler, L., (2016) Digital Materiality with Enhanced Reflectance Transformation Imaging. Archiving Conference. 19. April 2016 Vol., no. 1, p. 11-14

Mapping the Enlightenment: Intellectual Networks and the Making of Knowledge in the European Periphery

Vassilis Routsis
v.routsis@ucl.ac.uk
Centre for Digital Humanities
University College London, United Kingdom

Eirini Goudarouli
egoudarouli@gmail.com
The National Archives, United Kingdom

Manolis Patiniotis
mpatin@phs.uoa.gr
National and Kapodistrian University of Athens, Greece

The project *Mapping the Enlightenment: Intellectual Networks and the Making of Knowledge in the European Periphery* is funded by the Research Centre for Humanities (RCH) for the academic year 2016-2017. The project uses interactive mapping tools for visualising, exploring, and analysing the intellectual and geographical networks developed by Greek-speaking scholars of the Ottoman Empire during the 17th and 18th centuries. Based on the convergence of the latest achievements in digital mapping and historiographical discussions about the representation of the Enlightenment, the project develops user-friendly interactive and dynamic web maps of the itineraries of traveling scholars, and visually represent the building of networks between scientific centers and peripheries. Additionally, this dynamic system provides multi-layered maps that enables users to query and visualize data and flows through a modern and robust environment. This interactive interface offers a simple and effective way of showing how the intellectual networks developed within the European periphery in the 17th and the 18th centuries contributed to the shaping of knowledge during the Enlightenment. The project is a collaboration between the Department of History and Philosophy of Science of the National and Kapodistrian University of Athens and the Centre for Digital Humanities of University College, London.

The theoretical background of the project resides in the history of science. Following the general idea that the knowledge systems that gradually dominated European modernity have come into being through a dynamic and multi-layered process, the project is based on the notion of 'moving localities' introduced by historians involved with the Science and Technology in European Periphery (STEP) network. The concept of 'moving localities' enables historians to perceive circulation as a knowledge production process. Locality, in this sense, indicates a complex set of connections, allegiances and commitments which travel with people and thus extend beyond conventional boundaries, creating interconnected intellectual spaces over wide geographical areas. The sense of locality enables actors to perform distinct cultural identities in the course of their travels that are informed, but not confined by those assigned by their places of origin. Thus, locality can be said to be a local culture made active and open to transformation thanks to encounters fostered by travel conditions.

The project takes advantage of the theories of the STEP network on the one hand and the latest achievements in digital mapping on the other, in order to create a visual representation of the above-mentioned concept of 'moving localities'. The creation of a research tool will contribute to enhancing users' understanding of the emergence of modern science and technology as the expression of a dynamical geography. Addressing the spatiality of knowledge, the project focuses on associating particular cultural traits with specific points on a map, and work on tracking down the various paths and encounters through which such cultural traits and the respective knowledge practices evolved. In this context, centres and peripheries are not regarded as tokens of a steady, hierarchical geography, but rather as mutually dependent and co-constructed entities whose status can change with time.

The development of the user-friendly interactive interface initially focuses on the intellectual and geographical travels of the 18th-century Greek-speaking scholars, who developed an extended network connecting the most important European educational centers with the most important power centers of the Ottoman world. Adopting such an approach allows more historical actors to enter the scene, such as the Portuguese *estrangeirados* and Spanish *pensionados*. The visualisation of the paths followed by the Greek-speaking scholars, the *estrangeirados* and the *pensionados* enriches this historical and digital representation of the Enlightenment in the periphery.

The technical plan aims to use and supports open source software. The project ambitiously strives to make every possible bit of information queryable and visualised. On the server side, Apache, PostgreSQL, PHP and GeoServer with PostGIS library are the principal technologies that are used for the processing and serving of the data. On the client side, the latest versions of the web standards model HTML5, CSS3 and JavaScript provides a modern and user-friendly user interface. Initially, the scholars flow data are processed, normalized and ingested into the geo-enabled PostgreSQL database. The visualisations of the data are materialised using the Leaflet library and other JavaScript libraries such as D3.js, Chart.js etc. The combination of the above technologies gives life to our historical data by combining powerful visualisation components and a data-driven approach to DOM manipulation.

This project does not aim to create visuals and maps per se; instead, it perceives these techniques as the medium to assess and display the findings of historical research that critically challenges the received narrative concerning early modern intellectual European networks. The creation of a visual image of the dynamic, multi-layer process of knowledge-shaping in the European periphery challenges current digital projects which mainly reflect the mainstream positivist histories of the Enlightenment. The intellectual networks developed within and across the periphery are mainly absent from the mainstream histories of the Enlightenment and if they are recognized, they are typically treated as the paths of intellectually parochial scholars who were unable to fully embrace the ideal of modernisation through reason and science. Bringing such figures to the forefront, and confirming their role in the production of scientific and technical knowledge, may help historians tell more nuanced stories about the complex cultural encounters, which molded the European intellectual space and the multifarious knowledge exchanges that shaped the notion of European science and technology. By using modern digital techniques, we aim to enhance the picture of the European periphery as a historiographical standpoint in order to transcend the established spatial hierarchies and bring to the fore the continuous re-inventions, conceptual shifts and cultural adjustments, which are responsible for the shaping of modern scientific and technical knowledge. The proposed project presents an innovative narrative of the making of knowledge during the Enlightenment. At the same time, it constitutes a unique combination of the latest web and mapping technologies, and provides innovative means of storing, transferring, visualising, and querying historical data.

We believe that this project is a perfect example of how Digital Humanities, through its interdisciplinary nature of binding together research in humanities with digital technologies, can generate new critical knowledge through the re-interpretation of data that might otherwise be obscured. In addition, with the use of the latest open-source technology in spatial visualisation, the project adds value to the field by providing a showcase of the evolution of relevant technologies and, more importantly, the ways that Digital Humanities can innovatively use them for research. Finally, the project has the potential to offer a basis for a more ambitious project on a larger scale following the successful completion of its first stage. The ultimate aim is to allow other researchers to submit their data to the open-architected and forward-compatible queryable mapping tool and provide a historically more accurate overview of the intellectual movements during the Enlightenment.

Bibliography

Bodenhamer, D.J., Corrigan, J. and Harris, T.M. (*eds*). (2010). 'The Spatial Humanities: GIS and the Future of Humanities Scholarship'. Bloomington: Indiana University Press

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. and Schnapp, J. (2012). 'Digital Humanities'. MIT Press

Duncan, A.S. (2016). 'Online interactive thematic mapping: Applications and techniques for socio-economic research'. *Computers, Environment and Urban Systems*. vol. 57, pp. 106-117

Gavroglu, K., Patiniotis, M., Papanelopoulou, F., Simões, A., Carneiro, A., Diogo, M.P., Bertomeu-Sánchez, J.R., García, Belmar A. and Nieto-Galan, A. (2008). 'Science and Technology in the European Periphery: Some historiographical reflections'. *History of Science*, vol. xlvi, pp. 153-175

Presner, T. and Shepard, D. (2015). 'Mapping the Geospatial Turn'. In: Schreibman S., Siemens R. and Unsworth J. (*eds*). *A New Companion to Digital Humanities*. Chichester: John Wiley & Sons Ltd, pp. 199-212

Raposo, P.M.P., Simões. A., Patiniotis, M., Bertomeu-Sánchez, J.R. (2014). 'Moving Localities and Creative Circulation: Travels as Knowledge Production in 18th-Century Europe'. *Centaurus*, vol. 56, no. 3, pp. 167-188

Simões, A., Carneiro, A., Diogo, M.P. (*eds*). (2003). 'Travels of Learning. A Geography of Science in Europe'. Dordrecht: Kluwer Academic Publishers

Une ontologie pour archiver et donner accès aux œuvres numériques littéraires

Yan Rucar
yanrucar@hotmail.com
Labex Obvil, Université La Sorbonne, France

Jean-Gabriel Ganascia
jean-gabriel.ganascia@lip6.fr
Labex Obvil, Université La Sorbonne, France

Resumé

Nous avons conçu une ontologie destinée à indexer le corpus des œuvres numériques littéraires. Cette ontologie prend en compte les dimensions esthétiques et matérielles des œuvres de sorte que l'analyse critique et la production d'archives se croisent dans la tâche de conservation.

Abstract

Depuis le premier générateur de textes conçu par Theo Lutz en 1959, la littérature créée sur ordinateur n'a cessé de produire des œuvres marquées par le caractère éphémère des logiciels et des ordinateurs. Sitôt le support obsolète, l'objet se désintègre. Les traces laissées par l'œuvre sont le plus souvent frustrantes, à la façon d'une fenêtre à peine ouverte sur un paysage riche en textures, et bientôt close. Les textes critiques comprennent un bref descriptif de l'œuvre, augmenté parfois de quelques captures d'écran. Le lien entre les mots et les images, le dynamisme de dispositifs en constant mouvement, le fonctionnement du programme sont perdus. Jamais les exégèses ne font figurer le code-source, la vie de l'organisme fébrile de l'œuvre numérique est dès sa désagrégation insituable. Les traces laissées par les œuvres numériques sont aussi frustrantes que ces chefs d'œuvres du cinéma muet aujourd'hui seulement disponibles sous forme d'une poignée de photographies, d'un scénario et d'une composition musicale. Ce qui animait et faisait le film manque, et c'est du coup le cœur du dispositif qui n'est plus qu'une vaste lacune, un centre absent. Le parallèle avec le cinéma est pertinent dans la mesure où l'œuvre numérique est mobile, caractérisée par la transformation. La littérature électronique est déjà historiquement riche, tout en étant trouée d'oubli. La bibliothèque est constituée pour une bonne partie d'ouvrages fantômes.

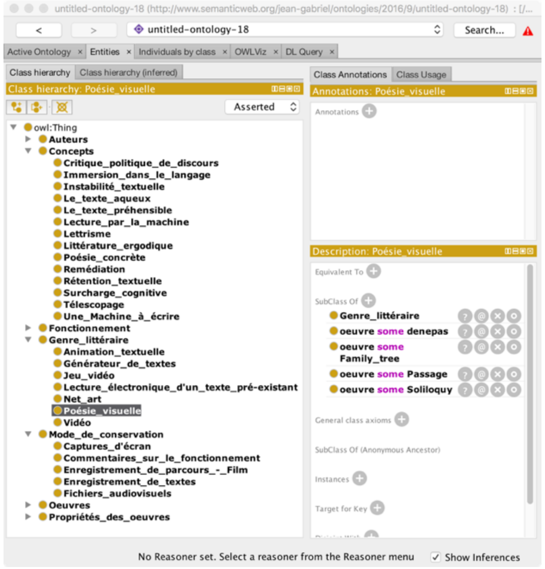
L'obsolescence des supports n'est pas le seul problème d'accès, l'incompatibilité des équipements est également un obstacle important à la lecture des œuvres. Ainsi les chefs d'œuvres de la fiction hypertextuelle *Afternoon, a story* (1987) de Michael Joyce et *Patchwork girl* (1995) de Shelley Jackson ne sont lisibles que sur des ordinateurs Mac. D'autre part, les différences entre ordinateurs peuvent complètement changer la perspective sur l'œuvre, en l'éloignant de l'intentionnalité de l'auteur.

Dès lors se fait jour la nécessité de concevoir un point de référence à ces œuvres changeantes et éphémères, un site électronique qui par-delà l'obsolescence prévisible des logiciels et l'absence d'interopérabilité des systèmes, puisse proposer des traces pertinentes de ces objets fuyants. En somme, la tâche à accomplir revient à archiver une œuvre existante, à la concevoir aujourd'hui même comme un présent révolu, à la considérer dans sa mortalité potentielle. Quand l'archivage est accompli après la désagrégation des œuvres, leurs vestiges consistent en des descriptions lacunaires dans des articles critiques, des images qui sont des fragments des objets numériques, parfois quelques mots sur les programmes. Ces traces manquent de pertinence car elles sont le fruit d'actes isolés, dispersés, et tout à fait fortuits. Des aspects essentiels des œuvres ne sont pas couverts par ces archives parcellaires. En créant l'archivage a priori paradoxal d'une œuvre encore accessible, on se rend maître des traces en cessant d'être soumis aux effets du hasard, on trie les faits saillants d'une esthétique au détriment de propriétés secondaires : « Plutôt que l'idée de conserver certains aspects, il faut se demander quels aspects abandonner » (Rouffineau 2016). Quand l'œuvre disparaîtra, ses traits définitoires seront encore consultables dans le site d'archives.
Un tel site aura plusieurs finalités :

- constituer une mémoire de l'objet numérique
- disposer des substrats pouvant servir de supports pédagogiques
- proposer une pluralité de points de vue sur une œuvre protéiforme
- contribuer à un discours critique visant à la définition d'un genre littéraire

En effet, les archives pourront être utilisées dans un cursus universitaire afin de pallier aux difficultés d'accès aux œuvres, et leur diversité corrigera les distorsions d'ordinateur à ordinateur, qui tendraient à une perspective faussée sur l'objet d'étude.

Dans cette perspective nous avons conçu une ontologie pour indexer les œuvres à partir de leurs différentes caractéristiques, qui constituent autant de points d'entrées aux œuvres littéraires numériques. L'ontologie a été construite à l'aide de l'éditeur Protégé (Kapoor et Sharma, 2010) et elle existe dans le format OWL. La photographie d'écran qui suit en donne une image partielle.



Class hierarchy: Poésie_visuelle

```
owl:Thing
  Auteurs
  Concepts
    Critique_politique_de_discours
    Immersion_dans_le_langage
    Instabilité_textuelle
    Le_texte_aqueux
    Le_texte_préhensible
    Lecture_par_la_machine
    Lettrisme
    Littérature_ergodique
    Poésie_concrète
    Remédiation
    Rétention_textuelle
    Surcharge_cognitive
    Télescopage
    Une_Machine_à_écrire
  Fonctionnement
  Genre_littéraire
    Animation_textuelle
    Générateur_de_textes
    Jeu_vidéo
    Lecture_électronique_d'un_texte_pré-existant
    Net_art
    Poésie_visuelle
    Vidéo
  Mode_de_conservation
    Captures_d'écran
    Commentaires_sur_le_fonctionnement
    Enregistrement_de_parcours_-_Film
    Enregistrement_de_textes
    Fichiers_audiovisuels
  Oeuvres
  Propriétés_des_oeuvres
```

Annotations: Poésie_visuelle

Description: Poésie_visuelle

```
SubClass Of
  Genre_littéraire
  oeuvre some denepas
  oeuvre some Family_tree
  oeuvre some Passage
  oeuvre some Soliloquy
```

No Reasoner set. Select a reasoner from the Reasoner menu

Cette ontologie comporte des descripteurs précisant les concepts de la littérature numérique, les fonctionnements des programmes informatiques associés, les genres littéraires, les modes de conservation possibles, et les propriétés multi-médiatiques des œuvres. Les descripteurs font à la fois état de l'esthétique et de la matérialité des œuvres. Le versant esthétique est représenté par les catégories des concepts et des genres littéraires, qui tentent de circonscrire l'intentionnalité des dispositifs. Le concept est le moteur, la justification de l'entité formelle. Bien que les descripteurs soient nombreux (voir tableau 1), on peut dégager trois familles conceptuelles, allant du complet abandon aux tourbillons de l'écran à la recherche du contrôle sur le texte numérique. La contemplation de l'action de la machine, la fascination technologique, représente le premier ensemble conceptuel, au sein duquel le spectateur suit du regard les découpages textuels opérés par le générateur. Le descripteur « immersion dans le langage » décrit bien une seconde tendance visant à entourer le lecteur/spectateur/internaute de mots à peine visibles, qui vont par exemple se télescoper, un mouvement dont l'effet est de faire perdre à l'utilisateur une vision surplombante par rapport au langage pour faire de cet instrument de sens un facteur de désorientation. La tendance inverse, celle de la recherche d'un contrôle, est bien résumée par le concept de littérature ergodique forgé par Espen Aarseth (1997). La compréhension d'un mécanisme complexe permet l'accès au texte. Cette dernière modalité correspond dans le tableau des genres littéraires au jeu vidéo (voir tableau 2). Les genres littéraires peuvent être classés en quatre regroupements : la poésie visuelle, l'animation, le générateur de textes et le net-art. La poésie visuelle et l'animation se fondent sur des effets immersifs dans le langage, l'interactivité accroît cette dimension. Le générateur de textes délègue à la machine la production littéraire, et le spectateur du net-art s'abandonne à un tissage des flux mis en scène par le programme.

Les descripteurs des fonctionnements (tableau 3), des modes de conservation (tableau 4) et des propriétés multi-médiatiques des œuvres (tableau 5) quittent le champ conceptuel pour s'intéresser à la matérialité des écrans. Les fonctionnements (tableau 3) sont génératifs (le programme produit du texte), interactifs, ou bien une séquence est déroulée (film vidéo, logiciel cartographique), enfin l'ordinateur s'ouvre aux réseaux (liens avec des journaux en ligne). Cinq principaux types de documents sont conservés (tableau 4): le texte, l'image, le son, le parcours sous forme d'enregistrement, le programme (code-source). La dimension textuelle des œuvres (tableau 5) s'articule autour des notions d'appropriation et de transformation, de fragmentation, de couplage entre le signifié et le signifiant graphique. La dimension sonore peut être centrale ou bien un arrière-plan, les descripteurs « musique liée au texte » et « reconnaissance vocale du texte » désignent les œuvres électroniques régulées par ou régulant des fichiers sonores. La dimension visuelle peut être inscrite dans l'apparence du signifiant graphique du texte, peut être autonome mais liée sous forme d'images évoluant au rythme des mots, ou bien l'iconicité est indépendante du langage.

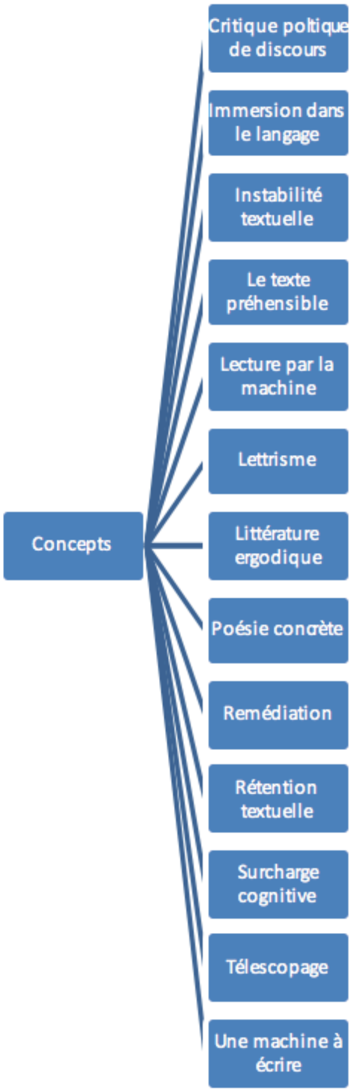


Tableau 1. Descripteurs de concepts

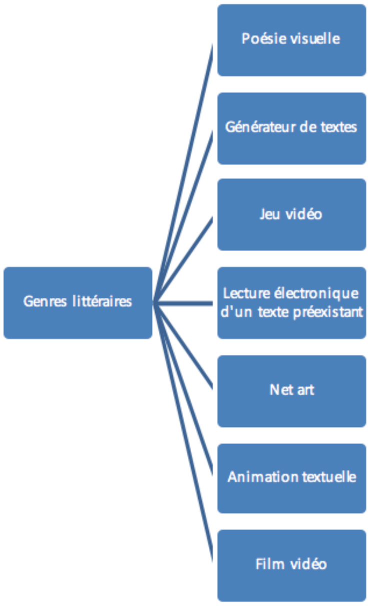


Tableau 2. Descripteurs des genres littéraires

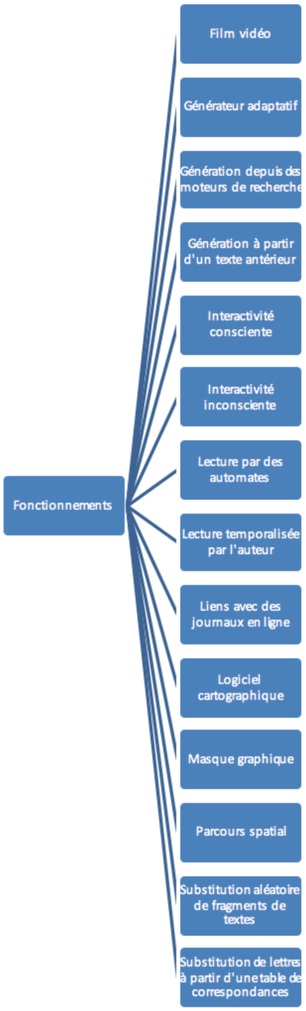

Fonctionnements
Film vidéo
Générateur adaptatif
Génération depuis des moteurs de recherche
Génération à partir d'un texte antérieur
Interactivité consciente
Interactivité inconsciente
Lecture par des automates
Lecture temporalisée par l'auteur
Liens avec des journaux en ligne
Logiciel cartographique
Masque graphique
Parcours spatial
Substitution aléatoire de fragments de textes
Substitution de lettres à partir d'une table de correspondances

Tableau 3. Descripteurs des fonctionnements

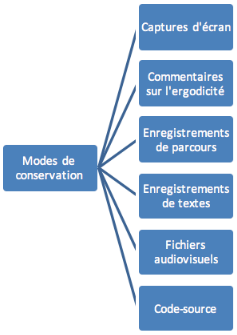

Modes de conservation
Captures d'écran
Commentaires sur l'ergodicité
Enregistrements de parcours
Enregistrements de textes
Fichiers audiovisuels
Code-source

Tableau 4. Descripteurs des modes de conservation

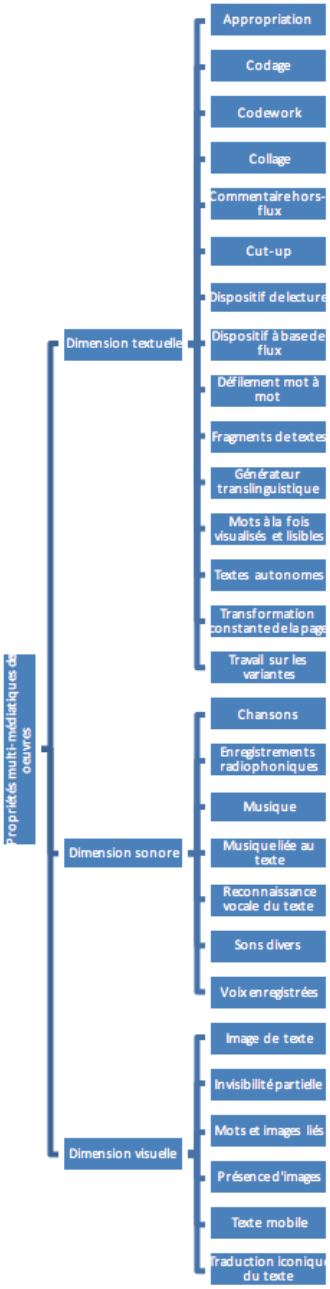


Tableau 5. Descripteurs des propriétés multi-médiatiques des œuvres

Cette ontologie est le premier pas vers la création de l'interface d'un site d'archives, dans lequel coexisteront des substrats, tels que décrits dans le tableau 4, et des points de vue critiques. Alexandra Saemmer souligne cette nécessaire mise en perspective : « La capture n'est pas suffisante, mais c'est une trace, un document, valorisable par une contextualisation » (Saemmer 2016). Parallèlement à l'étude critique des œuvres, qui situera la pertinence des documents, une recherche de toutes les expériences lectorales, spectatorales, perceptuelles, interactives, sera menée à travers le web. En raison des différences d'ordinateurs, il est nécessaire de puiser dans les nombreux blogs et forums de discussion toutes les descriptions de rencontres et d'interactions avec les œuvres numériques étudiées. En glanant et en organisant ces expériences distinctes voire divergentes, on rend compte de l'œuvre fuyante et protéiforme qu'est l'œuvre numérique, et on sauve de l'oubli des témoignages reposant sur des supports éminemment friables. Les postulats de ces œuvres éphémères dépendant de programmes, seront rendu pérennes par des documents figés qui seront autant d'artefacts de type muséal.

Bibliographie

Lutz ,T. (1959). Cette œuvre fondatrice a été reprogrammée et est disponible en ligne : http://auer.netzliteratur.net/0_lutz/lutz_original.html La conservation du programme du générateur originel a permis sa remise en fonction.

Joyce, M. (1987). *Afternoon, a story*, Watertown, US: Eastgate systems.

Jackson, S. (1995) *Patchwork girl*, Watertown, US: Eastgate systems, 1995

Rouffineau, G. (2016), « Le plaisir du paratexte », in Pamal, preservation & art – media archeology lab et Labex Arts-H2H. *Cette pièce est en cours de maintenance. Merci de votre compréhension – Préservation des écritures numériques,* Journée d'étude, Paris, 27 octobre 2016

Kapoor, B., Sharma, S. (2010), A comparative study ontology building tools for semantic web applications. International journal of Web & Semantic Technology (IJWesT), 1(3), 2010

Aarseth, E. J. (1997) *Cybertext : perspectives on ergodic literature*, Baltimore : Johns Hopkins University Press.

Saemmer, A. (2016), « Conclusion », in Pamal, preservation & art – media archeology lab et Labex Arts-H2H. *Cette pièce est en cours de maintenance. Merci de votre compréhension – Préservation des écritures numériques,* Journée d'étude, Paris, 27 octobre 2016

Distant Rhythm: Automatic Enjambment Detection on Four Centuries of Spanish Sonnets

Pablo Ruiz Fabo
pabloruizfabo@gmail.com
Lattice Lab, CNRS, France

Clara Martínez Cantón
cimartinez@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Thierry Poibeau
thierry.poibeau@ens.fr
Lattice Lab, CNRS, France

Introduction

Enjambment takes place when a syntactic unit is broken up across two lines of poetry (Domínguez Caparrós, 2000: 103), giving rise to different stylistic effects (e.g. increased emphasis on elements of the broken-up phrase, or contrast between those elements), or creating double interpretations for the enjambed lines (García-Paje, 1991).

In Spanish poetry, the syntactic configurations under which enjambment takes place have been described extensively, and detailed studies on the use of enjambment by individual authors exist (see Martínez Cantón, 2011 for an overview) including, among others Quilis (1964), Domínguez Caparrós, (2000), Paraíso, (2000), Spang (1983) for a description of enjambment, and Alarcos (1966), Senabre (1982), Luján (2006), Martínez Fernández (2010) for case-studies on a single author. However, a larger-scale study to identify enjambment across hundreds of authors spanning several centuries, enabling distant reading (Moretti, 2013), was not previously available.

Given that need, we have developed software, based on Natural Language Processing, that automatically identifies enjambment in Spanish, and applied it to a corpus of approx. 3750 sonnets by ca. 1000 authors, from the 15th to the 19th century. What is the interest of such large-scale automatic analyses of enjambment? First, the literature shows a debate about which specific syntactic units can be considered to trigger enjambment, if split across two lines, and whether lexical and syntactic criteria are sufficient to identify enjambment. Second, the stylistic effects that enjambment permits are also an object of current research (Martínez Fernández, 2010). Systematically collecting large amounts of enjambment examples provides helpful evidence to assess scholars' current claims, and may stimulate novel analyses. Finally, our study complements Navarro's (2016) automatic metrical analyses of Spanish Golden Age sonnets, by covering a wider period and focusing on enjambment.

The abstract is structured thus: First we provide the definition of enjambment adopted. Then, our corpus and system are described, followed by an evaluation of the system. Finally, findings on enjambment in our diachronic sonnet corpus are discussed. The project's website provides details omitted here for space reasons, including samples for the corpus, results, and other details.

Enjambment in Spanish

Syntactic and metrical units often match in poetry. However, this trend has been broken since antiquity for various reasons (Parry (1929) on Homer, or Flores Gómez (1988) on early classical poetry).

In Spanish tradition, enjambment (in Spanish, "encabalgamiento") is considered to take place when a pause suggested by poetic form (e.g. at the end of a line or across hemistichs) occurs between strongly connected lexical or syntactic units, triggering an unnatural cut between those units.

Quilis (1964) performed poetry reading experiments, proposing that the following strongly connected elements give rise to enjambment, should a poetic-form pause break them up:

- **Lexical enjambment:** Breaking up a word. We translated "lexical enjambment" from Quilis's terms "encabalgamiento léxico" or "tmesis".
- **Phrase-bounded enjambment**: Within a phrase, breaking up sequences like "noun + adjective", "verb + adverb", "auxiliary verb + main verb", among others. We translated "phrase-bounded enjambment" from "encabalgamiento sirremático".
- **Cross-clause enjambment:** Between a noun antecedent and the pronoun heading the relative clause that complements the antecedent. We translated "cross-clause enjambment" from Quilis's "encabalgamiento oracional".

The project site includes Quilis's complete list of syntactic environments that can trigger enjambment, as well as the types identified by our system. Besides the enjambment types above, Spang (1983) noted that if a subject or direct object and their related verbs occur in two different lines of poetry, this can also feel unusual for a reader, even if the effect is less pronounced than in the environments identified by Quilis. To differentiate these cases from enjambment proper, Spang calls these cases "enlace", translated here as "expansion".

Quilis (1964) was the only author so far to gather recitation-based experimental evidence on enjambment. His typology is still considered current, and was adopted by later authors, although complementary enjambment typologies have been proposed, as Martínez Cantón (2011) reviews. Our system identifies Quilis' types, besides Spang's expansion cases.

Corpus

The corpus is based on two public online collections from Biblioteca Virtual Cervantes (García González, R. (ed.), 2006a, 2006b). The first one covers 1088 sonnets by 477 authors from the 15th-17th centuries. The second one contains 2673 sonnets by 685 authors from the 19th century. We created scripts to download the poems, remove HTML and extract dates of birth and death for the authors (About 30% of the 15th to 17th century authors had exact dates of birth and death, for the rest only the centuries were available. Among the 19th century authors, ca. 45% had exact dates of birth and death). Table 1 shows the distribution of authors and poems by century. The corpus covers canonical as well as minor authors, inspired in distant reading approaches (Moretti, 2007, 2013).

Period *	Sonnet Count	Sonnet %	Author Count	Author %
14.5	43	1.14	2	0.17
15	2	0.05	2	0.17
15.5	8	0.21	5	0.43
16	141	3.75	58	4.99
16.5	411	10.93	108	9.29
17	478	12.71	300	25.82
17.5	5	0.13	2	0.17
18.5	13	0.35	6	0.52
19	1150	30.58	361	31.07
19.5	1510	40.15	318	27.37
Total	*3761*	*100*	*1162*	*100*

Table 1: Distribution of sonnets and authors per period.

* Exact dates of birth and death are available for a minority of authors; often only the century was provided in the corpus sources. Periods ending in ".5" cover authors who lived in two centuries. E.g. period "15.5" covers authors born in the 15th and deceased in the 16th century

System description

The system has three components: a preprocessing module to format input poems uniformly, an NLP pipeline, and the enjambment-detection module itself.

The NLP pipeline is IXA Pipes (Agerri et al., 2014). Its results for contemporary Spanish are competitive. Our system uses it to obtain part-of-speech tags, syntactic constituency (e.g. verb-phrase, noun-phrase) and syntactic dependencies (e.g. direct object).

The enjambment detection module is rule and dictionary-based, and exploits the information provided by the NLP pipeline. Rules (30 in total) of different characteristics identify enjambed lines, assigning them a type among a list of 12 types, based on the typology in Section 2 (the full list of types identified by the system is available on the project site).

- Some rules are very shallow and only take **parts of speech** into account.
- Some rules additionally exploit **constituency** info.
- Some rules use **dependency** information, e.g. to detect "subject / object / verb" relations.
- For any type of rule, **custom dictionaries** can restrict rule application to a set of terms. E.g. certain verbs govern arguments introduced by one specific preposition; we itemized these verbs and their prepositions in a dictionary, to complement information provided by the NLP pipeline or correct parsing errors.

Enjambment annotations are output in standoff format. Further details can be found on the project's site.

System evaluation and discussion

Test-corpus

To evaluate the system, we created two reference-sets (SonnetEvol and Cantos20th), manually annotating enjambment in them

1. **SonnetEvol**: 100 sonnets (1400 lines) from our diachronic sonnet corpus of ca. 3750 sonnets (Table 2). This test-set contains 260 pairs of enjambed lines (in other words, if there is an enjambment between lines 1 and 2, we consider that as "pair of enjambed lines" in the reference corpus).
2. **Cantos20th**: 1000 lines of 20th century poetry (Colinas, 1983), showing natural contemporary syntax. We identified 277 pairs of enjambed lines.

The distribution of enjambment types in the test-corpora is balanced (Table 2). The SonnetEvol diachronic test-corpus is balanced across periods (Table 3). It should be noted that balancing across periods does not apply to the Cantos20th test-corpus: it covers the 20th century only.

We annotated the Cantos20th corpus in order to assess the system's performance on contemporary Spanish with natural diction, compared to its behaviour with the SonnetEvol corpus, which includes some archaic constructions and often shows an elevated register.

For the evaluation reported here, each sonnet was annotated by a single annotator. Obtaining multiple annotators' input on the same sonnet to assess inter-annotator agreement (Artstein and Poesio, 2008) is part of our ongoing work.

	Test-Corpus			
	SonnetEvol		*Cantos20th*	
Enjambment Types *	**Count**	**%**	**Count**	**%**
Total Phrase-Bounded	*104*	*40.00*	*175*	*63.18*
adj_adv	2	0.77	1	0.36
adj_noun	29	11.15	54	19.49
adj_prep	14	5.38	11	3.97
adv_prep	0	0	3	1.08
noun_prep	39	15.00	85	30.69
relword	1	0.38	2	0.72
verb_adv	5	1.92	7	2.53
verb_cprep	9	3.46	2	0.72
verb_chain	5	1.92	10	3.61
Total Cross-Clause[12]	*23*	*8.85*	*31*	*11.19*
Total Expansions	*133*	*51.15*	*71*	*25.63*
dobj_verb	65	25.00	39	14.08
subj_verb	68	26.15	32	11.55
Total All Types	*260*	*100*	*277*	*100*

Table 2: Distribution of enjambment types in the manually annotated reference corpora, providing counts and each type's percentage of the total enjambments per corpus. Counts refer to pairs of enjambed lines.

*The project site includes a description of each enjambment type.

SonnetEvol Test-corpus	
Period **	**Sonnet Count**
Total 15th-17th	*72*
14.5	3
15	2
15.5	2
16	14
16.5	21
17	27
17.5	3
Total 19th	*28*
18.5	3
19	17
19.5	8
Total All Periods	*100*

Table 3: Distribution of sonnets by period in the manually annotated SonnetEvol corpus. The 16th, 17th and 19th centuries cover ca. 30% of the corpus each, and the 15th century covers ca. 10% of the sonnets
**Exact dates of birth and death are available for a minority of authors; often only the century was provided in the corpus sources. Periods ending in ".5" cover sonnets for authors who lived in two centuries. E.g. period "15.5" covers sonnets for authors born in the 15th and deceased in the 16th century

Enjambment-detection tasks evaluated

We defined two enjambment-detection tasks:

- **Span-match**: the positions of enjambed lines proposed by the system must match the positions in the reference corpus for a correct result to be counted.
- **Typed span-match**: for a correct result, both the positions and the enjambment type assigned by the system to those positions must match the reference.

System results and discussion

Precision, recall and F1 were obtained. The definitions for Precision (P), Recall (R) and F1 were the usual:

$$F1 = 2\frac{P \cdot R}{P + R}\ ; \quad P = \frac{\text{nbr. of correct outputs}}{\text{nbr. of system outputs}}\ ; \quad R = \frac{\text{nbr. of correct outputs}}{\text{nbr. of reference outputs}}$$

Table 4 provides overall results for both corpora. Table 5 provides the per-type results on the diachronic test-corpus (SonnetEvol). The project's site contains more detailed results (e.g. per-type results for the Cantos20th corpus, or breakdowns for SonnetEvol per period).

Corpus	Task	N	P	R	F1
SonnetEvol	*span-match*	260	74.18	87.64	80.35
	typed span-match		61.24	72.31	66.31
Cantos20th	*span-match*	277	84.01	89.17	86.51
	typed span-match		78.04	83.39	80.63

Table 4: Overall enjambment detection results. Number of test-items, Precision (P), Recall (R) and F1 in our two test-corpora, for the span-match and typed span-match enjambment detection tasks

Enjambment or Expansion Type *	N	P	R	F1
Phrase-Bounded (all types)	*104*	*66.19*	*88.46*	*75.72*
adj_adv	2	100	50.00	66.67
adj_noun	29	54.55	82.76	65.75
adj_prep	14	58.82	71.43	64.52
noun_prep	39	55.36	79.49	65.26
relword	1	100	100	100
verb_adv	5	50.00	100	66.67
verb_cprep	9	83.33	55.56	66.67
verb_chain	5	100	80.00	88.89
Cross-Clause[12]	*23*	*76.00*	*82.61*	*79.17*
Expansions (all types)	*133*	*61.54*	*66.17*	*63.77*
dobj_verb	65	60.00	69.23	64.29
subj_verb	68	63.24	63.24	63.24

Table 5: Enjambment detection results per type. On the SonnetEvol corpus. Number of items per type, Precision (P), Recall (R) and F1 on the typed span-match enjambment detection task.
* The types are described on our site

For untyped detection (span-match), F1 reaches 80% in the SonnetEvol corpus, whereas F1 for typed detection is 66.31%. For the contemporary Spanish corpus (Cantos20th), F1 is higher: 80.63% typed detection, 86.51% span-match. This reflects additional difficulties posed by archaic language and historical varieties for the NLP system whose outputs our enjambment detection relies on. Expansions get lower F1 than phrase-bounded types overall. But we do not think that the F1 difference between SonnetEvol and Cantos20th is due to the higher proportion of expansions in SonnetEvol (Table 2): Results per-type (see the evaluation page of the project's site) show that phrase-bounded enjambment detection is 10 points of F1 lower in SonnetEvol than in Cantos20th. Also, phrase-bounded enjambment results for the 15th-17th period (with more archaic language) are 10 points of F1 lower than in the 19th century.

A common source of error was hyperbaton: the displacement of phrases triggers constituency and dependency parsing errors. Prepositional phrase (PP) attachment also posed challenges: Verbal adjuncts get mistaken for PPs complementing nouns or adjectives. This is a common problem in syntactic parsing, even for contemporary languages (see Agirre et al, 2008, for English). For historical varieties, Stein's (2016) results for verbal adjuncts and prepositional complements in Old French also suggest the difficulties posed by prepositional phrases.

Creating a reparsing module to manage hyperbaton and improve PP attachment results may be fruitful future work.

Scholarly results and discussion

The system's goal is detecting enjambment to help literary research on the phenomenon, via providing systematic evidence for its analysis.

We consider our untyped enjambed-line detection results helpful, given an F1 of ca. 80% on the diachronic test-set. As an example application, we examined the distribution of enjambment according to position in the poem, particularly in positions across a verse-boundary (lines 4-5, 8-9 and 11-12). Comparing the results for the 15th-to-17th centuries vs. the 19th century (Table 6 and Figure 1), we see that enjambment across the tercets increases clearly in the 19th century, with a small increase of enjambment across the quatrains (lines 4-5) and across the octave-sestet divide (lines 8-9). Given the manageable data volume, we validated the counts for enjambment across a verse boundary (Table 6) manually (but not the more voluminous data for all other positions).

The value of the tool is helping perform such analyses on a large corpus. This opens the door for scholars to assess the literary relevance of the findings, and search for the best interpretation.

Enjambed line positions	Scholarly relevance	15th-17th cent.		19th cent.	
		Count	%	Count	%
4-5	across quatrains	2	0.07	19	0.26
8-9	across octave-sestet divide	2	0.07	12	0.16
11-12	across tercets	20	0.72	147	2.01

Table 6: Pairs of enjambed lines across verse boundaries in the 15th-17th vs. the 19th centuries: Counts of enjambed line-pairs and percentages over the total number of enjambed line-pairs for each period. An example of the types of analyses stimulated by automatic enjambment detection

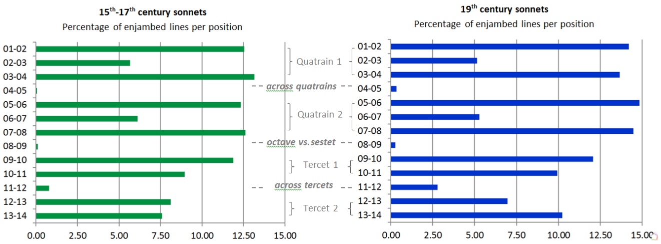


Figure 1: Percentage of enjambments per position in the 15th-17th centuries vs. the 19th.
The y-axis represents line-positions; the x-axis is the percentage of enjambed line-pairs for a position over all enjambed line-pairs in the period. Enjambment across quatrains and across the octave-sestet divide is very rare, with a small increase in the 19th century. The division between the tercets blurs in the 19th century, in the sense that enjambment across them is clearly higher than in the previous period

Outlook

The characterization of enjambment in Spanish literary theory has unclear points. Systematically obtaining enjambment examples is helping us find additional evidence to analyze these unclear points. Moreover, we are not aware of a systematic large-sample study of enjambment across periods, literary movements, or versification types in Spanish, or other languages. Automatic detection can help answer interesting questions in verse theory, which would benefit from a quantitative approach, complementing small-sample analyses. e.g.: To what an extent is enjambment used differently in free verse vs. traditional versification?

Students in our metrics classes are currently annotating enjambment for 450 sonnets. These annotations will permit inter-annotator agreement computation. We will also examine the possibility of using supervised machine learning to train a sequence labeling and classification model to complement our current detection rules.

Acknowledgements

Pablo Ruiz Fabo was supported by a PhD scholarship from Région Île-de-France. We also thank Clara Martínez Cantón's and Borja Navarro's metrics students for their ongoing sonnet annotation work, and Borja Navarro for introducing his students to the annotations required.

Bibliography

Agerri, R., Bermudez, J. and Rigau, G. (2014). IXA Pipeline: Efficient and ready to use multilingual NLP tools. In *Proceedings of LREC 2014, the 9th International Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Agirre, E., Baldwin, T. and Martinez, D. (2008). Improving Parsing and PP Attachment Performance with Sense Information. In *Proceedings of ACL 2008, Conference of the Association for Computational Linguistics*, 317-325. Columbus, Ohio, US.

Alarcos Llorach, E. (1966). *La Poesía de Blas de Otero [por] E. Alarcos Llorach*. Madrid, Anaya.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics 34.4*: 555-596.

Colinas, A. (1983). *Noche más allá de la noche*. Madrid, Visor Libros.

Domínguez Caparrós, J. (2000). *Métrica española.* UNED, Spain.

García González, R. (ed.) (2006a). *Sonetos del siglo XV al XVII*. Alicante, Biblioteca Virtual Miguel de Cervantes. Retrieved from http://www.cervantesvirtual.com/nd/ark:/59851/bmc2r439

García González, R. (ed.) (2006b). *Sonetos del siglo XIX*. Alicante, Biblioteca Virtual Miguel de Cervantes. Retrieved from http://www.cervantesvirtual.com/nd/ark:/59851/bmc4q861

García-Page, M. (1991) En torno al encabalgamiento. Pausa virtual y duplicidad de lecturas. *Revista de literatura* 53.105: 595-618.

Flores Gómez, M. E. (1988). Coincidencia y distorsión (encabalgamiento) de la unidad rítmica verso y las unidades sintácticas. *Estudios clásicos*, *30*(94): 23-42.

Luján Atienza, Á. L. (2006). *Desde las márgenes de un río: la poesía coral de Diego Jesús Jiménez*. Córdoba, Litopress.

Martínez Cantón, C. (2011). *Métrica y poética de Antonio Colinas* (PhD Dissertation from UNED, Spain). Sevilla, Padilla Libros.

Martínez Fernández, J. E. (2010): *La voz entrecortada de los versos.* Barcelona, Davinci Continental.

Moretti, F. (2007). *Graphs, Maps, Trees. Abstract Models for Literary History.* Verso.

Moretti, F. (2013). *Distant Reading.* Verso.

Navarro-Colorado, Borja, Lafoz, M. R. and Sánchez, N. (2016). Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. *Proceedings of LREC, Tenth International Conference on Language Resources and Evaluation:* 4630-4634. Portorož, Slovenia.

Paraíso, I. (2000). *La métrica española en su contexto románico.* Madrid, Arco Libros.

Parry, M. (1929). The distinctive character of enjambement in Homeric verse. In *Transactions and Proceedings of the American Philological Association* (60: 200-220). Johns Hopkins University Press, American Philological Association.

Quilis, A. (1964). *La estructura del encabalgamiento en la métrica* *española: Contribución a su estudio experimental.* Consejo Superior de Investigaciones Científicas.

Senabre, R. (1992). El encabalgamiento en la poesía de Fray Luis de León. *Revista de Filología Española, 62(1)*. Consejo Superior de Investigaciones Científicas.

Spang, K. (1983). *Ritmo y versificación. Teoría y práctica del análisis métrico.* Universidad de Murcia, Spain.

Stein, A. (2016). Old French dependency parsing: Results of two parsers analyzed from a linguistic point of view. In *Proceedings of LREC the 11th International Language Resources and Evaluation Conference*: 707-713. Portorož, Slovenia.

Tracing the Colors of Clothing in Paintings with Image Analysis

Cihan Sarı
cihan.sari@boun.edu.tr
Bogazici University, Turkey

Albert Ali Salah
salah@boun.edu.tr
Bogazici University, Turkey

Alkım Almıla Akdağ Salah
almilasalah@sehir.edu.tr
Istanbul Sehir University, Turkey

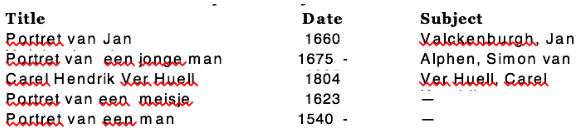
Introduction

The history of color is full of instances of how and why certain colors come to be associated with certain concepts, ideas, politics, status and power. Sometimes the connotations occur arbitrarily, like in the instance when pink was assigned to baby girls, and blue started to be associated with baby boys at the turn of 19th Century [Paoletti, 1987]. Sometimes though, color associations have very tangible reasons, such as in the case of Marian blue, reserved only for painting Virgin Mary over the centuries. The reason is found in the scarcity of the rock lapis lazuli –even more valuable than gold– from which the blue pigments were extracted. Individual colors have convoluted and contested histories, since they have been attached to many symbols at any given time. John Gage, an art historian who has devoted 30 years of research on the topic of color, explains the conundrum of what he terms "politics of color" in a simple fashion: "The same colors, or combinations of colors can, for example, be shown to have held quite antithetical connotations in different periods and cultures, and even at the same time and in the same place."[Gage, 1990].

The purpose of the present study is to introduce a method for automatically extracting color distributions and main colors of paintings, as well as color schemes of people in paintings. By visualizing these over time for cross-referencing with historical data, this study will reveal changes in how particular colors were used in a given time period and culture. In this study, we will look at artworks to find out whether certain colors or tones are associated with a specific sex, and if these connotations change over time. To that end, we apply algorithmic tools to process very large datasets automatically, and information visualization tools to depict the findings.

Related Work

Today, major cultural heritage collections are available online. Digitization and preservation of artworks is an important occupation of museums and cultural heritage institutions, as well as many Digital Humanities projects. Portions of of such digitized collections are made available to further computer vision research in order to scrutinize art historical questions. Such collections are usually enriched with meticulously tagged metadata describing the origins of each artwork. However, these datasets do not provide comprehensive gender annotations. For example, Rijksmuseum's arts database has a wide selection of categories with rich metadata that is primarily about the art objects themselves (see Table 1 – the quantity of meta information and context vary between different art samples), but without any reference to what these artworks hold [Mensink and Van Gemert, 2014]. Automatically determining whether a sitter of a portrait is female or male in a painting is not an easy task.



Title	Date	Subject
Portret van Jan	1660	Valckenburgh, Jan
Portret van een jonge man	1675 -	Alphen, Simon van
Carel Hendrik Ver Huell	1804	Ver Huell, Carel
Portret van een meisje	1623	—
Portret van een man	1540 -	—

Table 1: Sample from Rijksmuseum meta data

Several publications have appeared in recent years with the aim of automatic gender recognition. The survey by Ng et al. described a variety of approaches to gender recognition in natural images [Ng et al., 2012]. Xiong and De la Torre (2013) proposed a practical and effective method for automatically detecting faces in natural or man-made images. Once the face is detected, a supervised classifier is used to determine whether it belongs to a male or female. This requires the ground truth annotation of a large number of face images, from which the automatic classifier learns the visual boundary between these two classes.

There has been focused studies to address face recognition tasks on artistic images [Srinivasan et al., 2015]. For the purposes of face detection, mainstream algorithms perform sufficiently well on paintings that are of interest for this study. Automatic male/female classification is not perfect, it will occasionally get confused and produce an incorrect label. However, over thousands of images, a small number of individual errors will not prevent us from seeing the general patterns of color usage with males and females.

Methodology

In this study, the aim is to analyze the trends of clothing color in different periods, separately for males and females. For this purpose, we work on a database of paintings, for which the era (or date) is provided, and we seek to annotate each image with the gender of the depicted person, as well as a rough segmentation of the area of the clothing. The general workflow of the proposed approach is depicted in Figure 1.

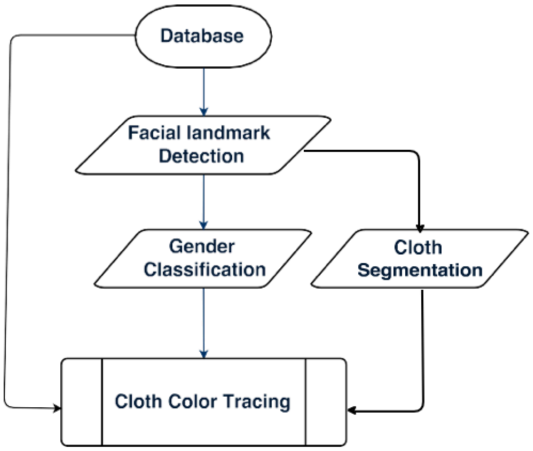


Figure 1. The workflow of the proposed approach.

Database

The Rijksmuseum is a Dutch national museum dedicated to arts and history in Amsterdam. The Rijksmuseum database contains 112.039 high-resolution im- ages with extended meta data [Mensink and Van Gemert, 2014]. However, as mentioned previously in Section 2, the Rijksmuseum database has neither gender nor clothing color information embedded into its metadata. We describe briefly how we determine the missing information.

Gender Classification

We have performed classification of the perceived sex from the face images. This process is commonly called Gender classification in computer vision – not to be mixed with characteristics of masculinity, femininity or sex organs, but what is perceived solely from the face crops on the paintings.

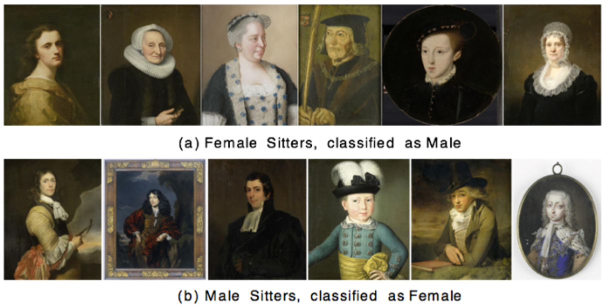
For this purpose we have prepared a test dataset of face images from Rijksmuseum paintings and three training datasets of face images: 10k US Adult Faces[Bainbridge et al., 2013], Labeled faces in the wild[Huang et al., 2007] and in an approach similar to Jia's work [Jia and Cristianini, 2015], we have gen- erated our own IMDB dataset. IMDB dataset images are collected using the Google Image search, using actor and actress names as queries. In total, 5600 male and 5300 female faces were downloaded.

None of the datasets have gender annotations, and hence we have performed face detection and facial landmark extraction methods in [Xiong and De la Torre, 2013], first, then hand-clean face detection and landmark extraction results against false positives and validate gender information (for all 10k US Adult Faces dataset and LFW dataset we had to manually annotate each image, but also Google Image search results for IMDB dataset are not perfectly robust, hence the IMDB dataset also had to be verified). Then we have aligned the faces to a mean shape [Gower, 1975], and extract features that are resistant to illumi- nation effects [Ojala et al., 2002]. We then train a classifier using the sequential minimal optimization (SMO) method [Platt et al., 1998].

The biggest challenge for evaluating gender recognition performance on the paintings is to make sure the ground-truth gender data are actually correct [Mathias et al., 2014]. From our experience, this demanding task requires a full view of the painting, rather than just the detected face. Results of some combinations of the datasets are given in Table 2. We could reach above 75% accuracy on paintings, just by using photographs of actors and actresses in the training of the system. Some of misclassification examples are given in figure 2.

	IMDB	IMDB and 10k	IMDB, 10k and LFW
Female	**62.16%**	62.32%	57.11%
Male	**84.51%**	83.40%	85.79%
Total	**77.21%**	76.41%	76.28%

Table 2. Gender recognition performance on Rijksmuseum. All results are com- parable and best (by small margin) is acquired when only the IMDB dataset is used.

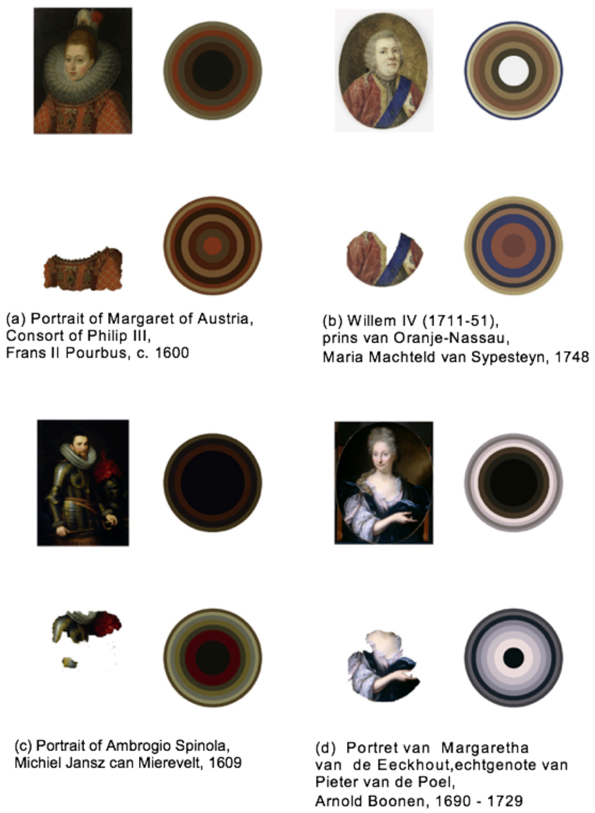

(a) Female Sitters, classified as Male

(b) Male Sitters, classified as Female

Figure 2. Misclassified paintings

Clothing color information

Portrait paintings that are completely focused on the sitter's face have still a lot of background noise that disrupt the color representation of the paintings (see Figure 3). Our hypothesis is that color representation, when focused on the clothing of the model, will still reflect the color scheme that is associated with the gender of the sitter.


(a) Portrait of Margaret of Austria, Consort of Philip III, Frans II Pourbus, c. 1600

(b) Willem IV (1711-51), prins van Oranje-Nassau, Maria Machteld van Sypesteyn, 1748

(c) Portrait of Ambrogio Spinola, Michiel Jansz can Mierevelt, 1609

(d) Portret van Margaretha van de Eeckhout,echtgenote van Pieter van de Poel, Arnold Boonen, 1690 - 1729

Figure 3. Four paintings from the Rijksmuseum collection, classified and segmented in terms of colors.

In order to extract the color information of an outfit we need to do a coarse segmentation of the clothing. We used the GrabCut approach [Rother et al., 2004]. In this method, a user defines an area of interest, as well as foreground and background seeds for the segmentation. In our study, background and foreground seeds are initialized based on the detected face landmarks.

Figure 4 provides an initial visualization of the dominant color distributions for each era, for males and females. Concentric circles have thickness associated with the frequency of the color. Bright colors are relatively rare, as the paintings in our tagged collection are generally dark, with the occasional shaft of light illuminating part of the painting. But a very distinct pattern can be observed in that females wear lighter colors compared to males, and show higher variance over the years. Some painting examples are given in Figure 5.

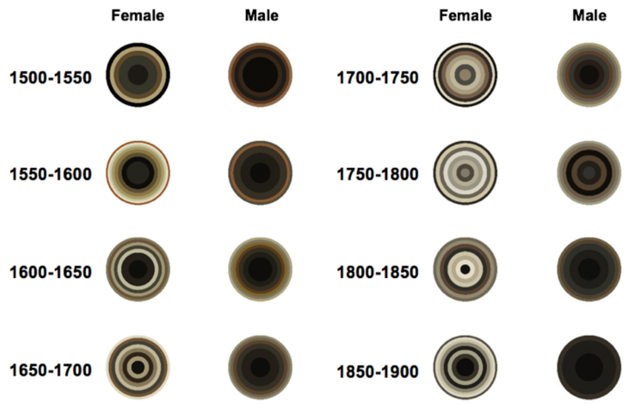
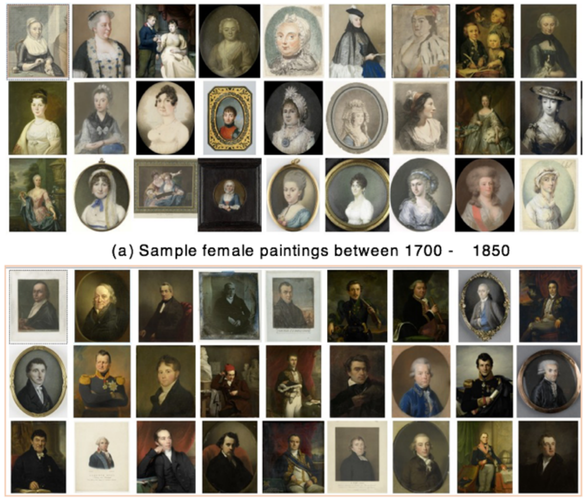


Figure 4. Clothing colors over time. Females wear significantly lighter colors than males. Best viewed in color.


(a) Sample female paintings between 1700 - 1850

(b) Sample male paintings between 1700 - 1850

Figure 5. Paintings of males and females from the Rijksmuseum database over time. Best viewed in color.

Conclusions

Every period and location has certain dominant color associations and symbolism. To investigate hundreds of thousands paintings in a single sweep requires automatic analysis tools. Our main objective in this work in progress is to perform an analysis on the usage of color for different genders along the centuries, and to develop tools for establishing semantic associations of colors for each particular period of study. With the increased popularity of open-art, this study can be extended significantly by introducing more databases alongside Rijksmuseum, for example, drawing on the Europeana collection [Doerr et al., 2010].

Bibliography

Bainbridge, W. A., Isola, P., and Oliva, A. (2013). The intrin- sic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4):1323.

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., and van de Sompel, H. (2010). The europeana data model (edm). In *World Library and Information Congress: 76th IFLA general conference and assembly*, pages 10–15.

Gage, J. (1990). Color in western art: An issue? *The Art Bulletin*, 72(4):518–541.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Jia, S. and Cristianini, N. (2015). Learning to classify gen- der from four million images. *Pattern Recognition Letters*, 58:35–41.

Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. (2014). Face detection without bells and whistles. In *Computer Vision–ECCV 2014*, pages 720–735. Springer.

Mensink, T. and Van Gemert, J. (2014). The ri- jksmuseum challenge: Museum-centered visual recognition. In *Proceedings of Inter- national Conference on Multimedia Retrieval*, page 451. ACM.

Ng, C. B., Tay, Y. H., and Goi, B.-M. (2012). Recognizing human gen- der in computer vision: a survey. In *PRICAI 2012: Trends in Artificial Intelligence*, pages 335–346. Springer.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Paoletti, J. B. (1987). Clothing and gender in America: Children's fashions, 1890-1920. *Signs*, 13(1):136–143.

Platt, J. et al. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Report MSR-TR-98-14, Microsoft Research*.

Rother, C., Kolmogorov, V., and Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM.

Srinivasan, R., Rudolph, C., and Roy-Chowdhury, A. K. (2015). Computerized face recognition in renaissance portrait art: A quantitative measure for identifying uncertain subjects in ancient portraits. *Signal Processing Magazine, IEEE*, 32(4):85–94.

Xiong, X. and De la Torre, F. (2013). Supervised de- scent method and its applications to face alignment. In *IEEE Conference on Com- puter Vision and Pattern Recognition (CVPR)*.

Interfacing Collaborative and Multiple-Layered Spaces of Interpretation in Humanities Research. The Case of Semantically-Enhanced Objective Hermeneutics

Christoph Schindler
schindler@dipf.de
German Institute for International Educational Research
Germany

Cornelia Veja
veja@dipf.de
German Institute for International Educational Research
Germany

Helge Kminek
kminek@em.uni-frankfurt.de
Goethe University Frankfurt, Germany

Introduction

In recent years, semantically enhanced Digital Humanities Research has become a widespread topic realized in different environments (e.g. CWRC, Pundit). While semantic graph technologies are mainly used to connect, annotate, query and aggregate strictly formalized entities, there is a lack of interfaces for enhancing acts of interpretations.

Annotations are described as crucial in interpretations and are designated as a killer application (Juola, 2009), scholarly primitive (Unsworth, 2000) and considered as notetaking within main scholarly information activities (Palmer, et al. 2009). Concerning an interpretational act, limitations of annotations are identified (overlapping, flexibility), and there is a demand for customization to the research context and an iterative and agile of schema development (Piez, 2010). Drucker indicates the emergent qualities of interpretation, while suggesting an interface which "supports acts of interpretation rather than simply returning selected results from a pre-existing data set" (Drucker, 2013: 37). In this paper, this desideratum is addressed by designing an interface for collaborative and multi-layered spaces of interpretations based on a semantic graph (Suchman, 2007; Drucker, 2011; Rheinberger, 2010; Barad, 2003).

A specific style of interpretation is chosen as a case study, i.e. collaborative analysis of face-to-face situations in small groups by creating a multiple layered space of interpretation - Objective Hermeneutics. In German-speaking countries, the approach of Objective Hermeneutics is one of the main methodologies used for qualitative analysis (Flick, 2005), which generates deep-structure analyses of cases by reconstructing actions and meanings. Creation of an interface that enables semantic annotations for these acts of interpretations makes it possible to elaborate and explicate a multiple layered space of interpretation.

In the following, we describe settings in the interpretational act of Objective Hermeneutics. Furthermore, the background of the design is outlined in relation to methods, design and data. The main contribution is twofold: (i) realization of an interface focusing the semantic enhancement of the collaborative spaces of interpretation and accountability and (ii) examines the semantic explication of the research data (interactional protocols), the interlinked multiple layered annotations and the possibility to retrace the space of interpretation.

The collaborative interpretational act of Objective Hermeneutics

The theoretical framework of Objective Hermeneutics is based on Oervermann's theory of professionalization (see Reichertz, 2004), whereby the act of interpretation follows strict principles for analyzing 'natural protocols' of social practices (transcripts). In a sequential multi-step procedure of interpretation, the structure of the case is reconstructed. The act of interpretation is realized in small groups where a common space of imagination is created collaboratively, wherein multiple layers of interpretations interfere and make use of falsification and abduction (Flick, 2005). Accordingly, the process of interpretation can be outlined as follows: 1) Specifying research question and analytical framework; 2) choosing appropriate transcript; 3) selecting sequence from interaction protocol (transcript); 4) creating and discussing step by step multiple corresponding stories, perspectives, and connections of the sequence; 5) recontextualisation to the concrete case, whereby in the long run hypotheses of the structure of the case are created iteratively and new sequences are selected (back to 3.). Additionally, 6) a proofing process is started (falsification). Based on this interpretative act, a detailed case structure is created which describes the conflicting motivations, interests and interactions of the actors. While in recent years special research data archives for archiving and re-using the transcripts (non-processable PDFs) have been established, the act of interpretation itself is still paper-based. This situation provides the opportunity for an appropriate case study for designing an interface for collaborative and multi-layered spaces of interpretations (for example, see the archive for pedagogical casuistry (ApaeK) archiving transcripts of classroom interactions)

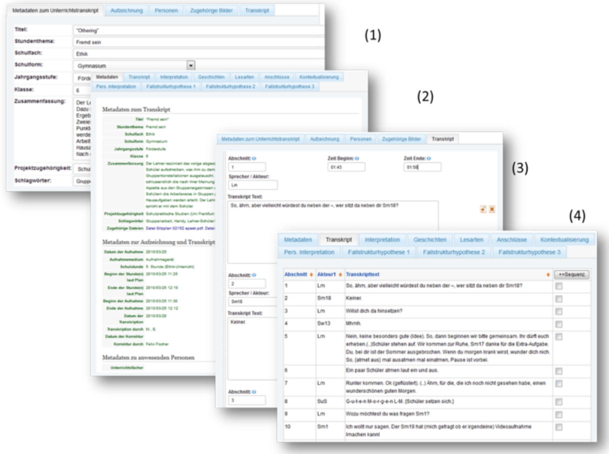
Methods, design, and data

The research environment for Objective Hermeneutics is based on a participatory design and agile development approach, using Semantic MediaWiki framework. To fulfill the case-related special research requirements an extension for Semantic MediaWiki and a research ontology were collaboratively developed. Besides the analysis of needs and requirements (site visit, artefact analysis) rapid prototyping was used and three versions of the environment were thoroughly tested (the logfile analysis between the last two versions indicates a clear improvement at the interpretation process by reducing the break up rate from 33% to 0%.) A group of distributed researchers across Germany and interested in classroom interactions (topic othering) used the environment in practice over several months and supported the design process by attending meetings and giving feedback.

Interfacing collaborative spaces of interpretations and accountability

Explicating interaction protocols semantically

The act of interpretation in Objective Hermeneutics is based on 'natural protocols' of social practice, which pursue strict notation guidelines for the transcription process (e.g. anonymization, settings of actors, properties like loudness). The transcript is enriched with contextual metadata (e.g. collecting context, duration, and topic). Line by line, each speech act of an actor and relevant interactions are described in detail (based on audio recordings, maps, photographs). This initial base already allows for semantically enhancing the space of interpretation: Interlinking relevant documents, entities, properties, and relations for semantic browsing (Figure 1, 1+2). Additionally, a formula semantically (Figure 1, 3+4) outlines the interactions of the transcript in detail (actor, speech act/interaction, line number). Thus, each annotation of interpretation can be related to this empirical level and the process of interpretation can be retraced. Based on this semantically enhanced transcript, the researchers choose and define their sequence of interest to start their act of interpretation (segment selection).



Titel: "Othering"
Stundenthema: Fremd sein
Schulfach: Ethik
Schulform: Gymnasium
Jahrgangsstufe: Förde
Klasse: 6
Zusammenfassung: Der L Dazu Ergeb Zweie Punkt werde Arbeit Hausa Nach
Projektzugehörigkeit: Schu
Schlagwörter: Grupp

Metadaten zum Transkript

Titel "Fremd sein"
Stundenthema Fremd sein
Schulfach Ethik
Schulform Gymnasium
Jahrgangsstufe Förderstufe
Klasse 6
Zusammenfassung Der Lehrer resümiert das vorige abges Schüler aufschreiben, was ihm zu dem Gruppenkonstellationen ausgetauscht. schlussendlich die nach ihrer Meinung Aspekte aus den Gruppenergebnissen Schülern die Arbeitsweise in Gruppen Hausaufgaben werden erteilt. Der Lehr spricht er mit dem Schüler
Projektzugehörigkeit Schulpraktische Studien (Uni Frankfurt
Schlagwörter Gruppenarbeit, Handy, Lehrer-Schüler
Zugehörige Dateien Datei:Sitzplan 02152 apaek.pdf, Datei

Metadaten zur Aufzeichnung und Transkript

Datum der Aufnahme 2010/03/25
Aufnahmemedium Aufnahmegerät
Schulstunde 5. Stunde (Ethik-Unterricht)
Beginn der Stunde(n) laut Plan 2010/03/25 11:25
Ende der Stunde(n) laut Plan 2010/03/25 12:10
Beginn der Aufnahme 2010/03/25 11:30
Ende der Aufnahme 2010/03/25 12:12
Datum der Transkription 2010/03/28
Transkription durch M., E.
Datum der Korrektur
Korrektur durch Felix Fischer

Metadaten zu anwesenden Personen

Unterrichtsfächer

Abschnitt: 1
Zeit Beginn: 01:43
Zeit Ende: 01:50
Sprecher / Akteur: Lm
Transkript Text: So, ähm, aber vielleicht würdest du neben der –, wer sitzt da neben dir Sm18?

Abschnitt: 2
Sprecher / Akteur: Sm18
Transkript Text: Keiner.

Abschnitt: 3

Abschnitt	Akteur1	Transkripttext	++Sequenz
1	Lm	So, ähm, aber vielleicht würdest du neben der –, wer sitzt da neben dir Sm18?	□
2	Sm18	Keiner.	□
3	Lm	Willst dich da hinsetzen?	□
4	Sw13	Mhmh.	□
5	Lm	Nein, keine besonders gute (Idee). So, dann beginnen wir bitte gemeinsam. Ihr dürft euch erheben.(..)Schüler stehen auf. Wir kommen zur Ruhe, Sm17 danke für die Extra-Aufgabe. Du, bei dir ist der Sommer ausgebrochen. Wenn du morgen krank wirst, wunder dich nich. So, (atmet aus) mal ausatmen mal einatmen, Pause ist vorbei.	□
6		Ein paar Schüler atmen laut ein und aus.	□
7	Lm	Runter kommen. Ok (geflüstert). (..) Ähm, für die, die ich noch nicht gesehen habe, einen wunderschönen guten Morgen.	□
8	SuS	G-u-t-e-n M-o-r-g-e-n L-M. (Schüler setzen sich.)	□
9	Lm	Wozu möchtest du was fragen Sm1?	□
10	Sm1	Ich wollt nur sagen. Der Sm19 hat (mich gefragt ob er irgendeine) Videoaufnahme machen kann!	□

Figure 1. Metadata of transcript (1, 2) and semantic interactions (3, 4)

Interlinking spaces of interpretation and multiple layered annotations

The act of interpretation in Objective Hermeneutics is semantically-enhanced and explicated by following guidelines for interpretation, whereby the flexibility of interpretation and the computer-mediated co-presence is taken into account. Each selected sequence of the transcript opens up the space of interpretation through multiple layered styles of annotations (stories, perspectives, connections, and contextualization) (Figure 2, 2+3). Subject to their discussions and notes, the researchers specify their arguments and elaborate a common ground for the case analysis. For an adequate interface of the phenomenon, the multiple layered styles of annotations need to be visualized in relation to the corresponding sequence of the transcript (compare Figure 2, 1+3). Closure respectively densification of the space of interpretation is semantically enhanced by creating specific case hypotheses. Researchers create connections between the layered annotations and specific case hypotheses, whereby a hypothesis and its related entities (e.g. layer of annotation, sequence, interaction, actor, or author) can be browsed semantically, described with texts and data representations. Each annotation is described with further relevant properties (e.g. timestamp, researcher name, related sequence) and interlinked within the sematic graph.

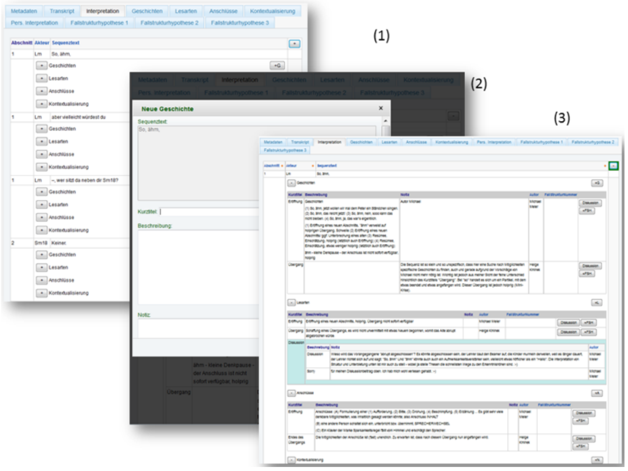


Figure 2. Selected sequences (1), annotation creation (2), and multiple layered annotations and discussions (3)

Retracing the spaces of interpretation and accountability

The analysis and reflection of the research process as well as the spaces of interpretation are enhanced by using a semantic graph and explicating the interaction protocols (transcripts) offering new capacities for retracing the spaces of interpretations (Figure 3). The multiple layered semantic graph interlinks the acts of interpretation and facilitates multiple perspectives for accountability concerning the 1) interaction protocols and their interlinking (Figure 4, 2), 2) the chronological acts of the researchers, 3) the multiple layered annotations for interpretations (Figure 4, 1), and 4) the creation and interlinking of the case hypothesis. Besides these imminent possibilities of the research project, external aspects of accountability can be addressed. While in Humanities the practice of data citations is not widely spread, research communities in Object Hermeneutics have established a citation practice via the archives of the transcripts, referring to the interactions of the transcript in their publications (as bibliographic data). In retracing the spaces of interpretation, the relevant multiple layers of annotations as well as the hypothesis interlinking can be referenced, opened, and described with semantic vocabularies. The semantic graph has been mapped to relevant semantic vocabularies e.g. Wf4Ever Research Object Model (ro), Object Reuse and Exchange (ore), Named Graphs (rdfg), Web Annotation Data Model.

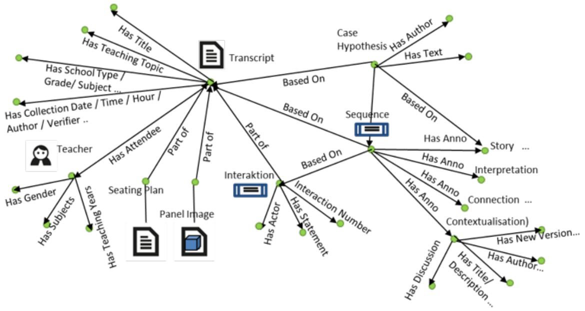


Figure 3. Visualization of semantic graph with traceable entities

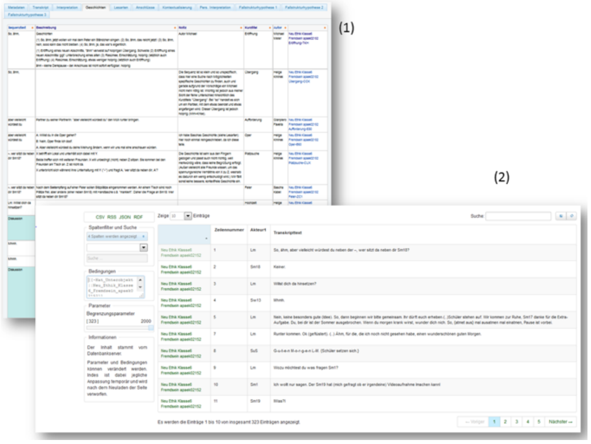


Figure 4. Annotation layer of stories (1) and interactive datatable of interactions (2)

Discussion and outlook

In this paper, we discussed and demonstrated the design of an interface for collaborative and multi-layered spaces of interpretation, using the methodological approach of ‘Objective Hermeneutics’ as a case study. The interface is considered in relation to the phenomenon and the relevant performative material-discursive capacities for the interpretational act, focusing on the use of a semantic graph. The detailed semantic description of the research data (transcripts) and the associated spaces of interpretations (stories, perspectives, connections, and contextualisations along with hypothesis) enable a collaborative and distributed analysis and new ways of retracing the spaces of interpretation (interlinked data, chronological acts, multiple layered annotations, case hypothesis). But time and effort of the semantic enhancement need to be balanced against these added values in each new research project (e.g. interlinking, accountability, data manipulation, visualisation, citation, and openness).

Acknowledgements

We would like to thank the research groups in Objectiv Hermeneutics and Semantic CorA, especially Marc Rittberger, Michael Meier and Sascha Kabel. Initially, the realization of the research environment Semantic CorA was supported by the German Research Foundation (DFG); its further development for Objective Hermeneutics was funded in the context of the eHumanities Center CEDIFOR by the German Federal Ministry of Education and Research (BMBF) no. 01UG1416C.

Bibliography

Barad, K. (2003). “Posthumanist performativity: Toward an understanding of how matter comes to matter.” Signs, 28(3), 801–831.

Drucker, J. (2013). “Performative Materiality and Theoretical Approaches to Interface.” digital humanities quarterly. 7 (1). http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html

Drucker, J. (2011). “Humanities approaches to interface theory.” Culture Machine. 12(0). 1-20.

Flick, U. (2005). “Qualitative Research in Sociology in Germany and the US—State of the Art, Differences and Developments.” Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6(3). http://www.qualitative-research.net/index.php/fqs/article/view/17

Juola, P. (2008). “Killer Applications in Digital Humanities.” Literary and Linguistic Computing, 23(1), 73–83. https://doi.org/10.1093/llc/fqm042

Palmer, C. L., Teffeau, L. C., Pirmann, C. M. (2009). “Scholarly information practices in the online environment.” Report commissioned by OCLC Research. http://www.oclc. org/programs/publications/reports/2009-02. pdf.

Piez, W. (2010). “Towards hermeneutic markup: an architectural outline.” Digital Humanities 2010: Conference Abstracts. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-743.pdf

Reichertz, J. (2004). “Objective Hermeneutics and Hermeneutic Sociology of Knowledge.” U. Flick, E. v. Kardoff, I. Steinke (eds.), A Companion to Qualitative Research. 290-295. London.

Rheinberger, H.-J. (2010). An epistemology of the concrete: Twentieth-century histories of life. Duke University Press.

Suchman, L. (2007). “Human-machine reconfigurations: Plans and situated actions.” Cambridge, University Press.

Unsworth, J. (2000). “Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this.” Symposium on Humanities Computing: Formal Methods, Experimental Practice. King’s College, London

The Third Way: Discovery Beyond Search and Browse in *Letters of 1916*

Susan Schreibman
susan.schreibman@nuim.ie
Maynooth University, Ireland

Sara Kerr
sarajkerr@icloud.com
Maynooth University, Ireland

Shane McGarry
shane.mcGarry@nuim.ie
Maynooth University, Ireland

Letters of 1916 is Ireland’s first digital public humanities project. Launched in 2013, it has become one of the key corpora in providing new insights and understandings of the 1916 period. The project collects and transcribes through crowdsourcing epistolary documents from October 1915 to November 1916 with the goal of creating a window onto a year in the life of the nation. In the middle of the project’s collection period is the Easter Rising (24-29 April 1916), arguably one of the most important events in Irish history as it sets in motion Irish independence from Great Britain in 1921.

The year 1916 was chosen not only because of the centrality of the Easter Rising, but because of Irish participation in the Great War and the historical significance for this throughout the following century in the construction of Irish identity. The collection also reestablishes the role of women in their participation in the Great War as well as the Rising and its aftermath. In addition, epistolary documents were chosen as a record of the everyday and the quotidian, providing a window onto a social history that has too often been repressed or ignored.

With a collection of over 3,500 letters contributed by 54 families and 32 institutions (with new letters being added continually) the corpus is too large to read in its entirety. While search and browse functionality in the project’s ‘Explore’ Database allows users to restrict results to more manageable subsets, the complexity of the letter form, with its frequently meandering content, makes many letters difficult to categorise using a tightly restricted set of keywords. Full text searching also misses many possible letters of interest on a particular topic due to the broad register and idiosyncratic use of language utilised by a wide variety of correspondents (Altman, 88).

A solution to these issues has been to explore the use of alternative methods of analysis and discovery, including topic modelling, vector space modelling with t-SNE vector reduction, and semantic network analysis. These methods provide alternative ways to explore a corpus of this size: too large to be read in its entirety via close reading, yet not big enough to qualify as big data. Rather, this type of collection, not out of reach of many DH projects, might be typified as one of the middle distance in which visualisations can serve as a series of lenses through which areas of interest can be identified for further research.

Topic modelling, although computationally expensive and requiring pre-processing, provides a detailed overview of the corpus and its themes. It highlights the complexity and variety of the letters’ content, as well as suggesting areas for further analysis. Combining the full text of the letters with the editorially-assigned keywords has proved a powerful combination in providing a bird’s-eye view of the corpus by theme.

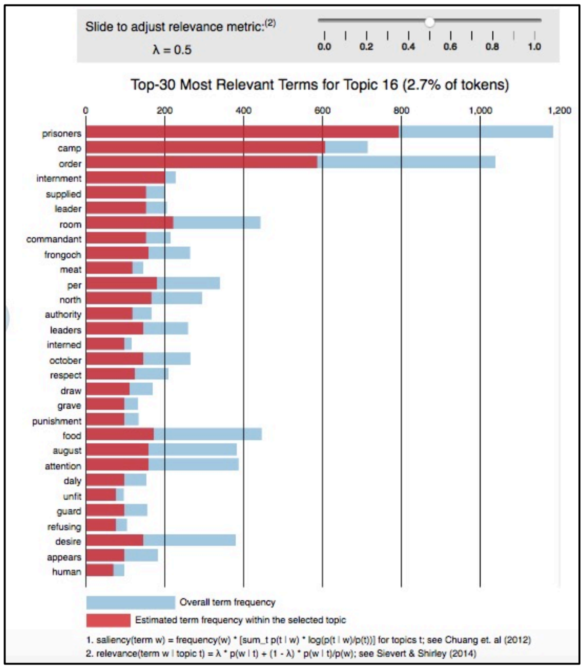


Figure 1. Murder at Portobello Barracks

Figures 1 and 2 show topics 14 and 16 from a model where only the standard stopwords were removed. Topic 14 relates to the murder of Sheehy Skeffington, while Topic 16 focuses on internment. The visualisation was created using 'LDAvis', where the most distinct terms in the topic can be viewed by adjusting the relevance metric λ to 0.5.

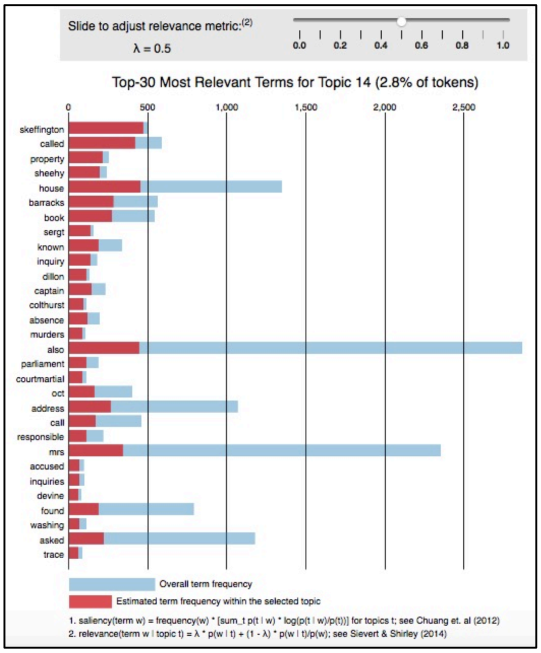


Figure 2. Topic 16 - Internment at Frongoch

Vector space modelling, using the R package 'wordVectors' (Schmidt and Li) which is based on the 'word2vec' algorithm (Mikolov et al.), is effective for a corpus where the researcher is familiar with the broad topics and wishes to zone in on specific aspects. Rather than asking 'which topics are in this corpus?', it allows the researcher to ask 'what does the corpus tell us about this topic?', revealing syntactic and semantic relationships. This type of analysis requires less pre-processing than topic modelling and does not exclude words with a low frequency.

The resulting vector space model can be interrogated, for example by extracting the words nearest to the word vector. For example, the vector for 'rising' (eg the Irish Rising) results in: 'outbreak', 'scene', 'leaders', 'theatre', and 'hostility'. While we may expect the term 'theatre' to refer to a theatre of war, close reading of the six references to 'theatre' in the letters reveals that two refer to an operating theatre while the remaining four are regarding theatre as a place of entertainment. Vector rejection can be used to exclude particular word meanings in cases of polysemy, thus allowing the researcher to specify which meaning they wish to search for. In the above example a researcher interested in entertainment could exclude the alternative meanings by rejecting those words which are related to theatre in the sense of 'hospital' and 'operation'. This would create a vector with the meaning 'theatre as a place of entertainment' enabling a more focused search.

The words nearest the desired search term can be visualised through vector space reduction to two dimensions. We use the Barnes-Hut implementation of t-SNE (t-distributed stochastic neighbour embedding). The visualisation of the 500 words nearest to the vector for 'rising' includes a cluster of words: 'portobello', 'murders', 'accused', 'colthurst' and 'dickson'. This refers to the murders of Sheehy Skeffington, Dickson and MacIntyre at Portobello Barracks on the orders of British officer Bowen-Colthurst.

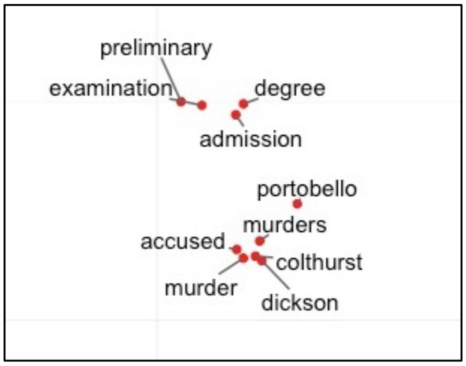


Figure 3. The 'Portobello' cluster

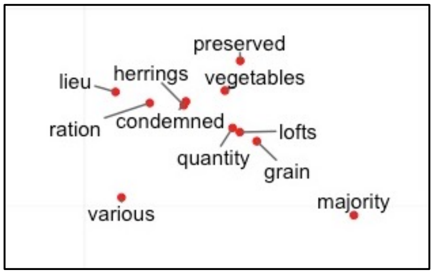
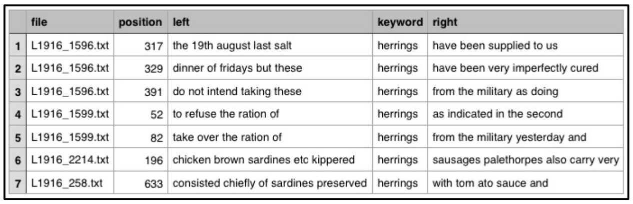


Figure 4. The 'herrings' cluster

A second cluster containing the words ‘herrings’, ‘condemned’ and ‘ration’ is less clear. A key word in context analysis demonstrates one of the strengths of the vector space model, the word ‘herrings’ only appears seven times in the corpus, but it sheds significant light onto the lives of those in internment. Figure 5 illustrates that the first five references to ‘‘herrings’ is in the context of the poor quality of rations provided to the Irish prisoners in Frongoch, a prisoner of war camp where some 3500 Irish men were sent after the British put down the Rising.



	file	position	left	keyword	right
1	L1916_1596.txt	317	the 19th august last salt	herrings	have been supplied to us
2	L1916_1596.txt	329	dinner of fridays but these	herrings	have been very imperfectly cured
3	L1916_1596.txt	391	do not intend taking these	herrings	from the military as doing
4	L1916_1599.txt	52	to refuse the ration of	herrings	as indicated in the second
5	L1916_1599.txt	82	take over the ration of	herrings	from the military yesterday and
6	L1916_2214.txt	196	chicken brown sardines etc kippered	herrings	sausages palethorpes also carry very
7	L1916_258.txt	633	consisted chiefly of sardines preserved	herrings	with tom ato sauce and

Figure 5. 'herrings' Key Word in Context

A third analysis builds upon the vector space model by creating a semantic network based on the cosine similarity of terms. The network is visualised using the ‘visNetwork’ R package (Almende and Thieurmel) which can be viewed as an interactive HTML file. This analysis enables the exploration of multiple dimensions of the model. Here terms can be clustered and connected with other terms providing a more detailed representation of the semantic space. This network can suggest a broader and more nuanced range of themes. Each of the analyses, and associated visualisations, provide a transformation of the text, disrupting expectations and providing new avenues for exploration (Clement).

These visualisations are, however, but the first step down the path of the third way. In order to engage the parahippocampal area of the brain, which is responsible for drawing context and meaning, we seek to create a more immersive experience for the reader (Bouchard et al). Thus a layer of interactivity is being explored to enhance these visualisations. While the visualisations themselves provide interesting insights into the corpus, they are, much like a traditional search and browse, rather static implementations in that they are pre-selected and curated. The project is thus exploring how users can be provided with the ability to customise these visualisations through the “slicing” of additional metadata in order to draw comparisons that may not be readily apparent.

However, this type of interactivity is not without its drawbacks. Both topic modelling and vector space visualisations require significant processing power in order to generate the base visualisations and the demands on internal memory of the processing machine are high as a result. Thus, creating these types of visualisations in an “on demand” environment is not feasible without a hardware investment that is beyond the reach of most of DH projects. Thus other options—such as caching, indexing, NoSQL databases, or various other data-related optimisation techniques—must be used in order to develop technical solutions that allow for interactivity while supporting a relatively low cost hardware solution.

These visualisations have begun to offer tantalising new insights into the corpus, providing a third way beyond search and browse. This paper will explore both the visualisations themselves, their strengths and weaknesses within the context of a corpus of the middle distance, as well as the novel readings they enable. The paper will conclude by discussing how interactive visualisations such as these can augment traditional modalities of interaction through a rich toolset for research and exploration.

Bibliography

Almende, B. V., and Thieurmel, B. (2016) *VisNetwork: Network Visualization Using “vis.js” Library*. https://CRAN.R-project.org/package=visNetwork. R package version 1.0.2.

Altman, J.G.. (1982) *Epistolarity: Approaches to a Form.* Columbus: Ohio State University Press. 1982.

Bouchard, S., et al. (2015) “The Meaning of Being There is Related to a Specific Activation in the Brain Located in the Parahypocampus.” 12th Annual International Workshop on Presence. November 2009. PDF. 11 July 2015.

Clement, T. (2013) “Text Analysis, Data Mining and Visualisations in Literary Scholarship.” *MLA Commons | Literary Studies in the Digital Age*, Oct. 2013, https://dlsanthology.commons.mla.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/

Mikolov, Tomas et al. “Efficient Estimation of Word Representations in Vector Space.” Proceedings of the International Conference on Learning Representations (ICLR 2013) (2013): 1–12.

Schmidt, B., and Li, J. (2015). *WordVectors: Tools for Creating and Analyzing Vector-Space Models of Texts.* R package version 1.3.

Integrating historical scientific texts into the Bernoulli–Euler Online platform

Tobias Schweizer
t.schweizer@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Sepideh Alassi
sepideh.alassi@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Martin Mattmüller
martin.mattmueller@unibas.ch
Bernoulli Euler Centre, University of Basel, Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Helmut Harbrecht
helmut.harbrecht@unibas.ch
Bernoulli Euler Centre, University of Basel, Switzerland

Introduction

Bernoulli-Euler Online (BEOL) is an interdisciplinary research project funded by the Swiss National Science Foundation focusing on the mathematics influenced by the Bernoulli dynasty and Leonhard Euler. It is being carried out by the Bernoulli Euler Centre and the Digital Humanities Lab at the University of Basel. Its main goal is the integration of different edition projects relating to the Bernoullis and Leonhard Euler into *one* target platform, offering appropriate functionality for researchers interested in the history of science.

The methodological efforts will also be applicable to other editions since they are developed in a generic way. BEOL is based on Knora, a generic infrastructure for humanities data.

Goal of the BEOL–platform and its technical basis

BEOL aims at integrating three edition projects, that are currently all technically different and thus incompatible with one another:

- *Basler Edition der Bernoulli-Briefwechsel* (BEBB): BEBB is an online edition that is based on the MediaWiki software and hosted by the University Library of Basel. It is connected to the library's metadata catalogue for manuscripts (Basler Inventar der Bernoulli-Briefwechsel). The letters are encoded in Wiki markup and are converted to HTML to represent them on the web. The mathematical formulae are encoded in LaTeX.
- *Leonhardi Euleri Opera Omnia* (LEOO): LEOO is a printed edition of the works of Leonhard Euler that was begun in the early 20th century. In the context of BEOL, the volume containing Euler's correspondence with Christian Goldbach (Euler 2015) will be integrated as a proof of concept. This volume has been prepared using LaTeX (as well as the volume with Euler's correspondence with Daniel Bernoulli that has been published recently). We expect to be able to integrate all the other recent volumes set in LaTeX in a similar manner. For the older volumes, the printed books would have to be scanned (including OCR) and marked up.
- Jacob (I) Bernoulli's scientific notebook *Meditationes*: The manuscript is held in the university library of Basel (shelfmark L Ia 3, 367 pages) and has already been digitized. The manuscript consisting of 287 entries is being transcribed at the Bernoulli Euler Centre using XML (The XML format is specified closely to the TEI specifications P5, so it can be transformed quite easily to TEI/XML) for the text and LaTeX for the mathematical notation that is embedded in the XML.

The three edition projects do not only overlap thematically, but also in terms of the persons involved (authors, mentioned persons) and bibliographical items (literature referred to in the texts, references in-between the editions' texts). Letters exchanged between members of the Bernoulli dynasty, Leonhard Euler and contemporary mathematicians and scientists are an important part of these edition projects and thus it is desirable to identify and match the persons in all editions in order to display their relations.

The technical basis for BEOL is Knora, an infrastructure for humanities data (Rosenthaler and others 2015) consisting of an RDF-triplestore, an OWL base ontology, and a RESTful API that allows for querying and adding to the data. The base ontology (see prefix 'Knora' in Figure 1) defines common value types (such as a calendar independent format to represent dates using the Julian Day Number) used among humanities projects and can be further extended in project specific ontologies. BEOL will provide such an ontology (see prefix 'BEOL' in Figure 1), defining its own resource classes and properties needed to represent the edition projects' texts and entities. Wherever possible, existing ontologies will be reused by making subclasses and subproperties. BEOL is part of the NIE-INE project, which aims to create a general-purpose infrastructure for digital editions, using Knora as its technical foundation. A focus of this project will be abstracting out concepts shared by different projects and formalising them as ontologies.

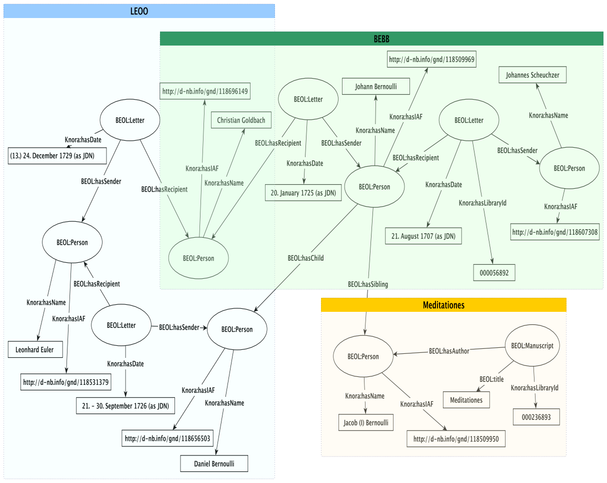


Figure 1: BEOL network and its components

Figure 1 represents all relations between persons (We refer to the Integrated Authority File (GND), and in order to represent locations, we will also refer to GeoNames), letters, and manuscripts (we also link to the catalogue of the Basel university library that keeps many of the original copies of the letters and manuscripts of BEOL), as well as their properties as directed graphs. For reasons of clarity, we use a simplified model here. The coloured rectangles indicate that these have been imported from different edition projects which – considered in isolation – do not allow for this kind of overview. Moreover, indices and bibliographies have to be unified on the BEOL platform (e.g., Christian Goldbach occurs both in BEBB and LEOO). The BEOL platform will be connected to Early Modern Letters Online, so it will be interoperable with other edition projects.

Importing editions to the same target environment

In order to represent all three editions in the same target environment, they have to be homogenised first. We decided to do so using an XML-based approach. This has the additional advantage that we can make both the texts of BEBB and LEOO available as TEI/XML to the outside world quite easily by applying XSL transformations. We can also use the same routine to import the editions into BEOL. Knora converts XML-encoded texts to RDF in order to store them in the triplestore. From RDF, an XML document can be recreated that is equivalent to the one originally imported. A mapping defines the relations between XML elements and attributes and the entities defined in the ontology.

- BEBB Wiki markup can be transformed to XML using a MediaWiki parser. Wiki tags and structures are mapped to XML tags, and references to other letters, bibliographical items, and images (facsimiles of the letters) can be handled. Once the letters are available on the BEOL platform, the old URLs will have to be forwarded.
- The Goldbach-volume of LEOO is set in LaTeX and can be converted to XML using LaTeXML. Additional mappings to the available standard functionality and customisations can be provided using Perl scripts. LaTeXML provides a MathML conversion for mathematical formulae.
- The Meditationes are transcribed in an XML-based format (see LaTeXML). Derived texts of these files can be generated using XSL-transformations. In this way, several layers (diplomatic, normalized) of the text can be produced. Our approach addresses segments defined on the facsimile (see Figure 2) and turns them into a reading text step by step. The figures (see segment 'M151-03-F' in Figure 2) will be extracted by applying a combination of various image processing techniques and redrawn as vector graphics.

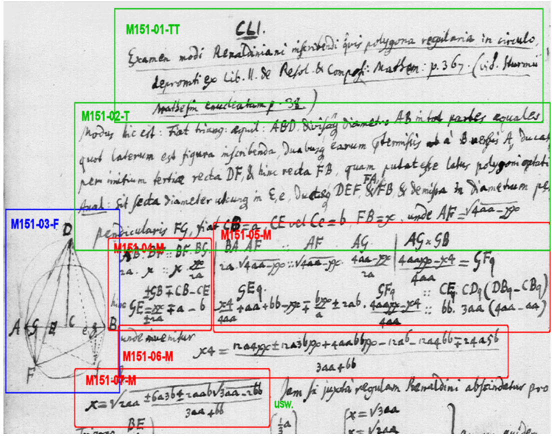

M151-01-TT
M151-02-T
M151-03-F
M151-04-M
M151-05-M
M151-06-M
M151-07-M
usw.

Figure 2. Part of Meditatio 151

One of the main challenges in the BEOL project is the faithful representation of mathematical notation and its relation to the surrounding text (see Figure 2) using web technologies. At the moment, we are using MathJax (which accepts both LaTeX and MathML as input formats) to render the mathematical formulae in the web browser. We also consider MathML as an option, although not all web browsers fully support MathML.

We are aiming at developing a browser based user interface that will be based on the Angular 2 framework (although in the meantime, we are already using Knora's current interface, SALSAH) that not only makes it possible to present the texts on the web and to offer search functionality, but also to add to the data (sufficient permissions provided). The users may create their own annotations on the BEOL platform. Basically, the user interface interacts with the Knora API in order to create new resources, manipulate properties etc. Since BEOL is based on Knora, all of its generic functionality can be used for this purpose.

Conclusion

BEOL integrates three different edition projects into one platform and allows researchers to query previously separated contents and add to them. The specific problems posed by the combination of text and mathematical notation can be addressed in a generic manner. All the functionality to be developed will be part of Knora and can be reused by other projects dealing with scientific texts from mathematics and physics.

Bibliography

Euler, L. (2015): *Leonhardi Euleri Opera Omnia*. Vols. IVA/4: Correspondence of Leonhard Euler with Christian Goldbach, ed. by Martin Mattmüller and Franz Lemmermeyer. Basel 2015.

Rosenthaler, L., et al (2015) *Final Report for the Pilot Project "Data and Service Center for the Humanities"*. Swiss Academy of Humanities and Social Sciences. http://www.sagw.ch/dms/sagw/laufende_projekte/DaSCH/FinalReport-DaSCH_print

Tracking transmission of details in paintings

Benoit Seguin
benoit.seguin@epfl.ch
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Isabella di Lenardo
isabella.dilenardo@epfl.ch
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
Digital Humanities Laboratory
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Introduction

In previous articles (di Lenardo et al, 2016; Seguin et al, 2016), we explored how efficient visual search engines operating not on the basis of textual metadata but directly through visual queries, could fundamentally change the navigation in large databases of work of arts. In the present work, we extended our search engine in order to be able to search not only for global similarity between paintings, but also for matching details. This feature is of crucial importance for retrieving the visual genealogy of a painting, as it is often the case that one composition simply reuses a few elements of other works. For instance, some workshops of the 16th century had repertoires of specific characters (a peasant smoking a pipe, a couple of dancing, etc.) and anatomical parts (head poses, hands, etc.) ,that they reused in many compositions (van den Brink, 2001; Tagliaferro et al, 2009). In some cases it is possible to track the circulation of these visual patterns over long spatial and temporal migrations, as they are progressively copied by several generations of painters. Identifying these links permits to reconstruct the production context of a painting, and the connections between workshops and artists. In addition, it permits a fine-grained study of taste evolution in the history of collections, following specific motives successfully reused in a large number of paintings.

Tracking these graphical replicators is challenging as they can vary in texture and medium. For instance, a particular character or a head pose of a painting may have been copied from a drawing, an engraving or a tapestry. It is therefore important that the search for matching details still detects visual reuse even across such different media and styles. In the rest of the paper, we describe the matching method and discuss some results obtained using this approach.

Method

Matching patterns in a bank of images is a problem that has been extensively studied in the Computer Vision community as « Visual instance retrieval » (Sivic and Zisserman, 2003). The definition of the task is : *given a region of a query image, can we identify matching regions in other images of a large iconographic collection* ?

Historically, the most successful methods were based on feature point descriptors (like SIFT, see Lowe, 2004) used in a Bag-of-Words (Jegou et al, 2008) fashion. The global architecture can be summarized as follows:

For each image in the collection:

- Extract the feature points for the image.
- Quantize the point descriptors to a Visual-Bag-of-Words representation.

Given a query region:

- Use the Bag-of-Words signatures to rank the first N most likely candidates of the collection.
- For each candidate, re-rank them according to a spatial verification of the matching points with the query region.

In practice, such an approach works extremely well and numerous improvements (Shen et al, 2009; Crowley and Zisserman, 2014) have been brought over the years to the different steps of the procedure. Hence it is still considered state of the art for the traditional datasets which focus on building and object retrieval in photographs.

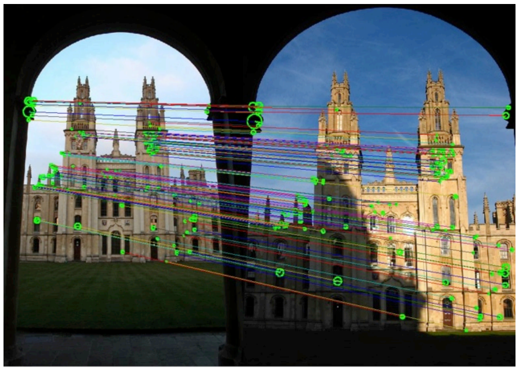


Figure 1: Working Bag-of-Words matching for buildings.

However, it was shown recently (Seguin et al, 2016; Crowley and Zisserman, 2014) that such approaches break completely when not dealing with the same physical objects and important style variations like we do in paintings. An example can be seen on Figure 2.

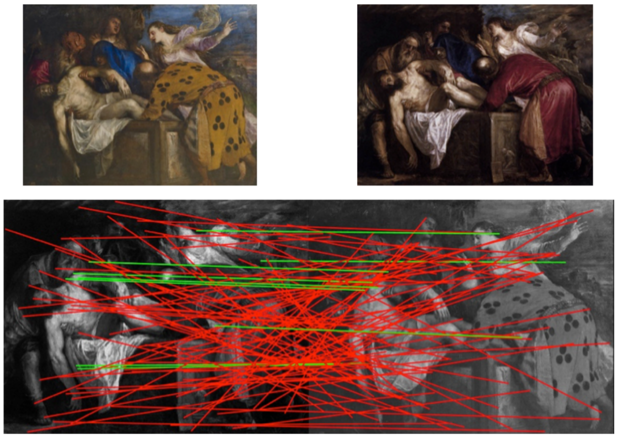


Figure 2: Feature point matching breaks when local features and style vary.

More recently, Convolutional Neural Networks (CNN) have had tremendous success in almost all areas of Computer Vision (object detection, recognition, segmentation, face identification, etc.) and CNN have established themselves over the last couple of years as an extremely powerful tool for almost any vision based problem.

A CNN is a multi-layer architecture where each layer transforms its input according to some parameters (also called weights). What makes them so powerful is that all these parameters can be learned « end-to-end » (for example in the case of object classification, just with images and their corresponding labels).

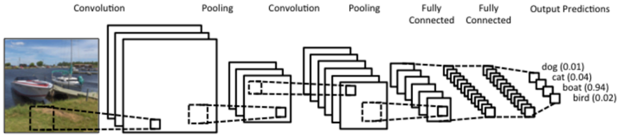


Figure 3: Simplified structure of a Convolutional Neural Network.

It has been shown in Donahue et al (2014) that CNN pre-trained on very large datasets, like Deng et al (2009), for object classification tasks offer a very good abstract representation of the image information, and are thus applicable for other vision problems. They generalize much better when transferring from photographs to paintings, contrary to the traditional Bag-of-Words techniques (Seguin et al, 2016; Crowley and Zisserman, 2014). However, its application to visual instance retrieval was always hindered by the fact that they traditionally output a single global descriptor for the image, hence not directly allowing for region (sub-image) retrieval (Babenko et al, 2014).

In Razavian et al (2014), the authors proposed to just precompute the CNN descriptors for some subdivision of the image. However, such approach limits the possible granularity of the windows, and multiply the memory requirement by a huge factor.

Another more promising approach introduced in Tolias et al (2015) is to work directly on the CNN feature maps. More precisely, a common way of extracting a CNN descriptor for image retrieval is to take an image of size *(H, W, 3)* (height of *H* pixels, width of *W* pixels and 3 color channels RGB) go through all the convolutional layers of CNN, which outputs the *feature maps* : a structure of size *(H', W', F)* (*F* channels of size *H'* and *W'*). From these feature maps, taking the sum (or the max) of each channel gives a signature of size F which is (after normalization) the descriptor of the image.

Traditionally, the network used is the VGG16 (Simonyan and Zisserman, 2014) architecture which given an image of size *(H, W, 3)* creates feature maps of size *(H/32, W/32, 512).*

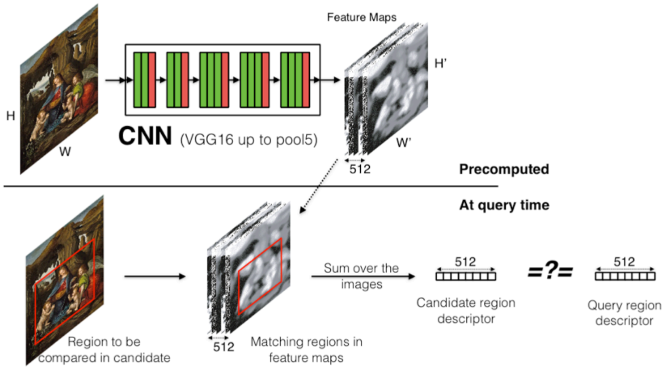


Figure 4: Region evaluation architecture.

Now, starting from the feature maps, computing the signature of a sub-part of an image is already easier, we just need to compute the sum of the corresponding region in the feature maps to obtain the descriptor of size *F*. However, evaluating many different regions would still be prohibitive from a computation point of view.

In order to alleviate the performance problem, Tolias et al (2015) proposed to use *integral images*. Given an image *I*, the integral image $I_\int$ is $I_\int(y,x) = \sum_{i<x,j<y} I(j,i)$. This allow for extremely quick computation of the sum of an image for a given area (y_1, y_2, x_1, x_2) (Fig.5) :

$$\sum_{x_1 \leq i < x_2, y_1 \leq j < y_2} I(j,i) = I_\int(y_2,x_2) + I_\int(y_1,x_1) - I_\int(y_1,x_2) - I_\int(y_2,x_1)$$

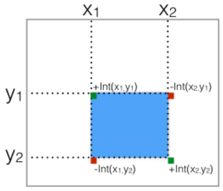


Figure 5: Integral images

This trick allows for extensive evaluation for the best matching window for the query image in the target collection. The global procedure for searching is the following, quite close to the Bag-of-Word approach:

For each image in the collection:

- Extract the feature maps of the image, and compute the corresponding integral images.
- Compute the global signature (with the whole image as window).

Given a query region:

- Use the global signatures to rank the first N most likely candidates of the collection.
- For each candidate, re-rank them according to an extensive look of the sub-windows in the images using their pre-computed integral images.

In order to greatly improve the results, we add the following improvements:

- The parameters of the network we use were fine-tuned using the Replica dataset (Seguin et al, 2016) (image retrieval in paintings). This dramatically improves the system resilience to color and style.
- We use Spatial Pooling according to Razavian et al (2014) which consists of extracting 4 blocks per evaluated region instead of 1. It makes the search roughly 4 times slower but allow for much better retrieval of complex patterns by directly encoding spatiality.

Results

The following experiment was run on the whole Web Gallery of Art collection (38'000 elements). Each image was resized so that its smaller dimension is 512 pixels, and the integral images of the feature maps computed on it. Given a query region for an image, the 300 most likely candidates are extracted from the WGA collection, and re-ranked according to the best matching window on each of them. Using 35 cores on a server machine, the complete request takes less than 4 seconds.

Examples of queries and their results are shown on Figure 6. In query 1 and 2, the starting image is Leonardo da Vinci's *Virgine delle Rocce* (Paris, Louvres), first version of this subject (1483-1486). Leornado is an interesting case study and his influence in Europe is known to be extremely important for the general compositions of paintings, character typologies and landscape patterns. In query 1, the group of Mary, the angel and the two children is selected. The first result is the version of the same subject (London, National Gallery) finished a decade after the Paris version, with the contribute of Ambrogio de Predis. The painting is different in color but has the same composition. The second and third results are other versions of the same subject by unidentified painters of the XVIth century. This constitutes typical examples of the propagation of a complex theme.

In query 2, only a detail from the landscape is chosen. Interestingly, the second result is a painting from Bernardino de Conti, which is a variation on the same theme, reusing the landscape but without the angel and with the two children kissing.

For query 3, we use a painting by Marco d'Oggiono, a follower of Leonardo, very similar to the one by de Conti and we select only the two children kissing. Results 1 and 3 feature paintings where only the children are present, showing that this replicator has an autonomy of its own. The third result by Joos van Cleve confirms the historical migration of this subject, as autonomous, from Italy to Flanders.

These 3 simple queries illustrate how the detail matching method can easily unveil the transmission network between different series of paintings.

Perspectives

The search of matching details in large-scale databases of paintings may enable to find undocumented links and therefore new historical connections between paintings. By tracking the propagation and transformations of a replicator, it becomes possible to follow the evolution through time of repertoires of forms and view each painting as a temporary vehicle playing the role of an intermediary node in a long history of images transmissions. Although in continuation with traditional methods in Art History, such a tool opens the avenue for research at a much larger scale, searching for patterns and finding new links simultaneously in millions of digitized paintings.

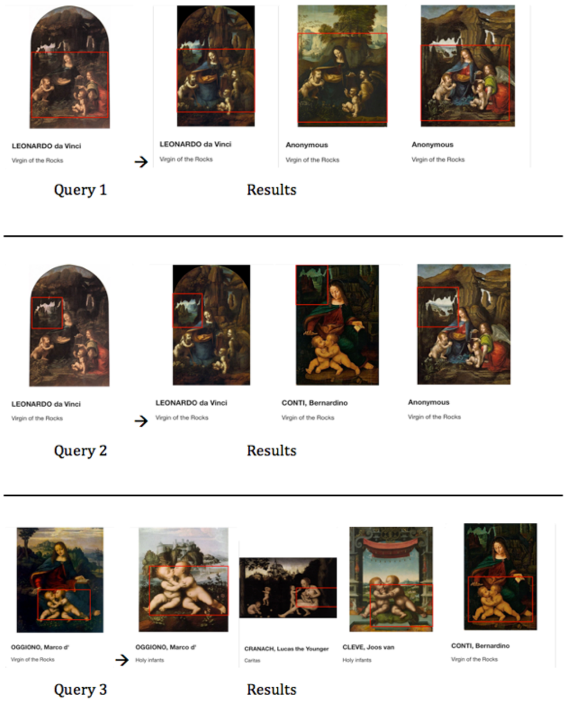


Figure 6: Examples of results of detail search.

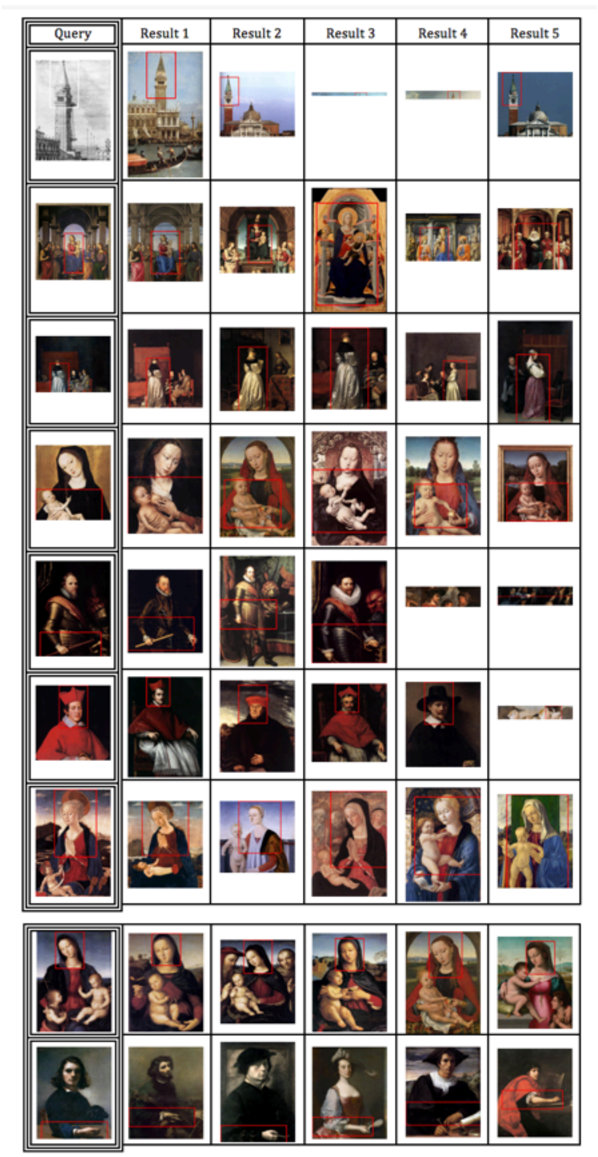


Figure 7: Additional examples of results of detail search.

Bibliography

Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V. (2014) “Neural codes for image retrieval,” in *ECCV*.

Crowley, E. J., and Zisserman, A. (2014) “In search of art,” *ECCV Workshops*, 2014.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014)“DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition,” *ICML*.

Deng, J., Dong, W., Socher, R. Li, L.-J., Li, K., and Fei-Fei, L. (2009) “ImageNet: A large-scale hierarchical image database,” *CVPR*.

Jegou, H., Douze, M., and Schmid, C. (2008) “Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search ”, *ECCV*.

di Lenardo, I., Seguin, B. L. A., and Kaplan, F. (2016). Visual Patterns Discovery in Large Databases of Paintings. Digital Humanities 2016, Krakow, Polland, July 11-16, 2016.

Lowe, D. G. (2004) “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vis.*, Nov. 2004.

Razavian, A. S., Sullivan, J., Maki, A., and Carlsson, S. (2014) “A Baseline for Visual Instance Retrieval with Deep Convolutional Networks,” Dec.

Seguin, B., Striolo, C., di Lenardo, I., and Kaplan, F. (2016) Visual link retrieval in a database of paintings, VISART : Where Computer Vision Meets Art, 3rd Workshop on Computer Vision for Art Analysis, October 2016, Amsterdam, The Netherlands

Shen, X., Lin, Z., Brandt, J., Avidan, S., and Wu, Y. (2012) “Object retrieval and localization with spatially-constrained similarity measure and k -NN re-ranking,” *CVPR*.

Simonyan, K., and Zisserman, A. (2014) “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv Prepr.*

Sivic, J., and Zisserman, A. (2003) “{Video Google:} A text retrieval approach to object matching in videos,” *CVPR*.

Tagliaferro, G., Aikema, B., Mancini, M., Martin, A. J.. (2009) *Le Botteghe di Tiziano Alinari*, Firenze

Tolias, G., Sicre, R., and Jégou, H. (2015) “Particular object retrieval with integral max-pooling of CNN activations,” *arXiv Prepr. arXiv1511.05879*

van den Brink, H. M., ed. (2001) L’entreprise Brueghel, , Maastricht, Bonnefantenmuseum- Bruxelles, Musée royaux des beaux-arts, Beaux-Arts Collection 2001

Informing Library-Based Digital Publishing: A Survey of Scholars’ Needs in a Contemporary Publishing Environment

Megan Senseney
mbonn@illinois.edu
University of Illinois, United States of America

LaTesha Velez
lmvelez2@illinois.edu
University of Illinois, United States of America

Christopher R. Maden
crism@illinois.edu
University of Illinois, United States of America

Janet Swatscheno
jswatsc2@illinois.edu
University of Illinois, United States of America

Maria Bonn
author.email@domain.com
University of Illinois, United States of America

Harriett Green
green19@illinois.edu
University of Illinois, United States of America

Katrina Fenlon
kfenlon2@illinois.edu
University of Illinois, United States of America

Introduction

When access as a value of scholarship is foregrounded in publishing, libraries emerge as "natural and efficient loci for scholarly publication" (Courant and Jones, 2015). In a rapidly evolving digital publishing landscape, academic libraries are poised to address scholars' publishing concerns about gaining access to opportunities for support and re-skilling, providing open access to their intellectual content, and ensuring access to the audiences who will most benefit from their work. The growth of library-based publishing services is evidenced by the 115 college and university libraries currently listed in the Library Publishing Directory (Lippincott, 2016). This paper presents selected results from a US-based survey on the needs of humanities scholars in a contemporary publishing environment, emphasizing aspects of the survey responses that shed light on the question of access in publishing from three perspectives: access to support services, access to content, and access to audience.

About PWW

Publishing Without Walls (PWW) is a Mellon-funded initiative at the University of Illinois led by the University Library in partnership with the School of Information Sciences, the department of African American Studies, and the Illinois Program for Research in the Humanities. Our project is developing an innovative and experimental library-based digital scholarly publishing model that aims to be accessible, scalable, and sustainable. Our objective is to develop a model for library-based publishing services that can be adopted broadly by other academic libraries to address scholars' emerging needs in a contemporary publishing environment. The model itself places humanities scholar at the center of the ecosystem, with services informed by—and responsive to—scholars' needs. Research and development within the project are strategically designed to address known gaps within the current landscape: the gap between what and how scholars want to publish and what existing systems of print publishing can accommodate; the gap between the everyday practices of humanities scholars and what high-level tools exist for producing and supporting digital scholarship; and the gap between digital scholarship and publishing opportunities at resource-rich institutions and Historically Black Colleges and Universities (HBCUs).

Survey Method

This paper presents the selected results of a large-scale survey about scholars' publishing practices and perceived needs. The full survey aims to understand what and how scholars want to publish, when and why they choose to publish digitally, and how they perceive the success of their digital publications. Survey outcomes will directly inform the development of PWW's shareable service model, but we also anticipate that our survey results will be relevant to digital humanists and other scholars engaged in digital scholarly publishing, whether such efforts are located within or beyond an academic library.

From June to October 2016, we conducted a large-scale survey of scholars, especially targeted at humanities scholars and scholars at HBCUs in the United States. The survey was developed by the PWW Research Team in spring 2016, and comprised around 30 questions. The survey covers six broad themes: respondents' experiences with print and digital publishing; respondents' publishing objectives; publishing tools and platforms; publishing services and support; publishing from the scholars' perspective of reader as opposed to author; and general attitudes toward print and digital publishing. The survey was distributed through listservs and social media venues targeting scholars in the humanities generally as well as selected niche communities to encourage sufficient responses across disciplines and institutions. The survey received 250 responses.

The team used the Qualtrics platform to present the survey and conduct initial analysis, with further quantitative data analysis in SPSS. Preliminary findings have been reported previously (Fenlon et al., 2016), and analysis is ongoing. The survey instrument and a summary report will also be archived in the IDEALS repository (Velez et al., in preparation).

Results

Respondents were asked to identify which aspects of publishing posed the most significant challenges with respect to their experiences with print and digital content. Figure 1 represents the percentage of respondents who indicated each potential issue as either "quite challenging" or "extremely challenging" in print or digital publishing.

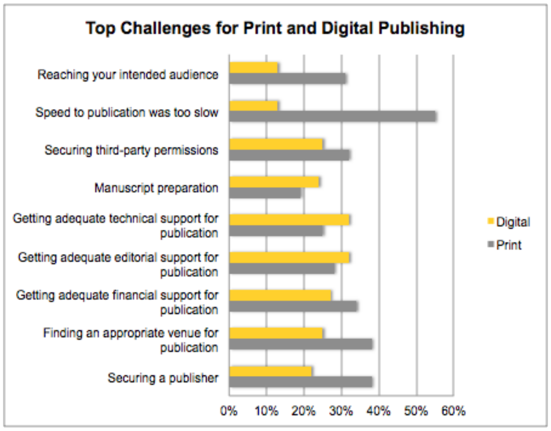


Figure 1. Top Challenges for Print and Digital Publishing

The top three challenges for digital publishing include getting adequate technical, editorial, and financial support for publication. Respondents were also asked to indicate and rank their top five publishing goals, which are illustrated in Figure 2 as a weighted bar graph where a first-place ranking is assigned 10 points, a second-place ranking is assigned 5 points, and a third-place ranking is assigned 1 point. The top three goals are consistent with the traditional expectations for scholarly publishing: contributing new information to one's field, encouraging and participating in dialogue about an area of study, and establishing a formal record of one's scholarship.

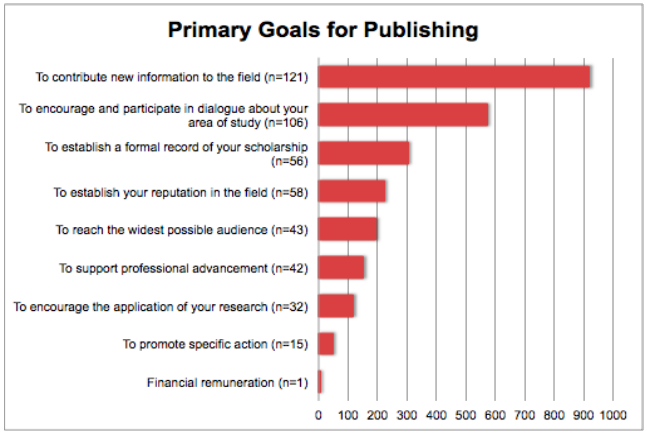


Figure 2. Primary Goals for Publishing

Figure 3 illustrates the top three audiences that scholars indicated they most wish to reach. The top two audiences relate to peers within the academic community. While interest in reaching the general educated reader and students is less frequently cited, it is sufficiently robust to consider how reaching these audiences may have an impact on scholars' decision making with regard to the medium they choose and the venues they seek for publication. Understanding how less traditional, but still prevalent, publishing goals affect these choices is also a potentially fruitful avenue for exploring how explicitly stated publishing objectives inform, and possibly shift, priorities regarding representation and dissemination within scholarly publishing. These themes are explored in a set of four charts in Figure 4.

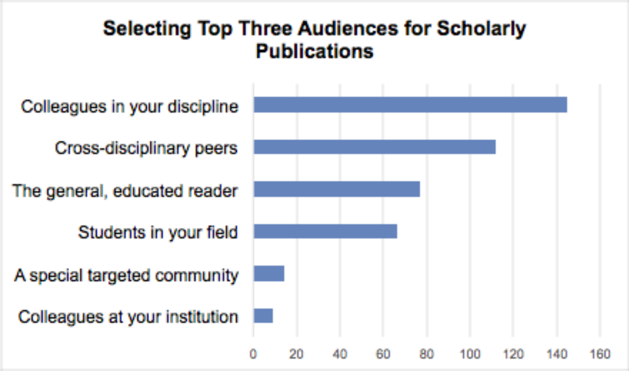


Figure 3. Selecting Top Three Audiences for Scholarly Publications

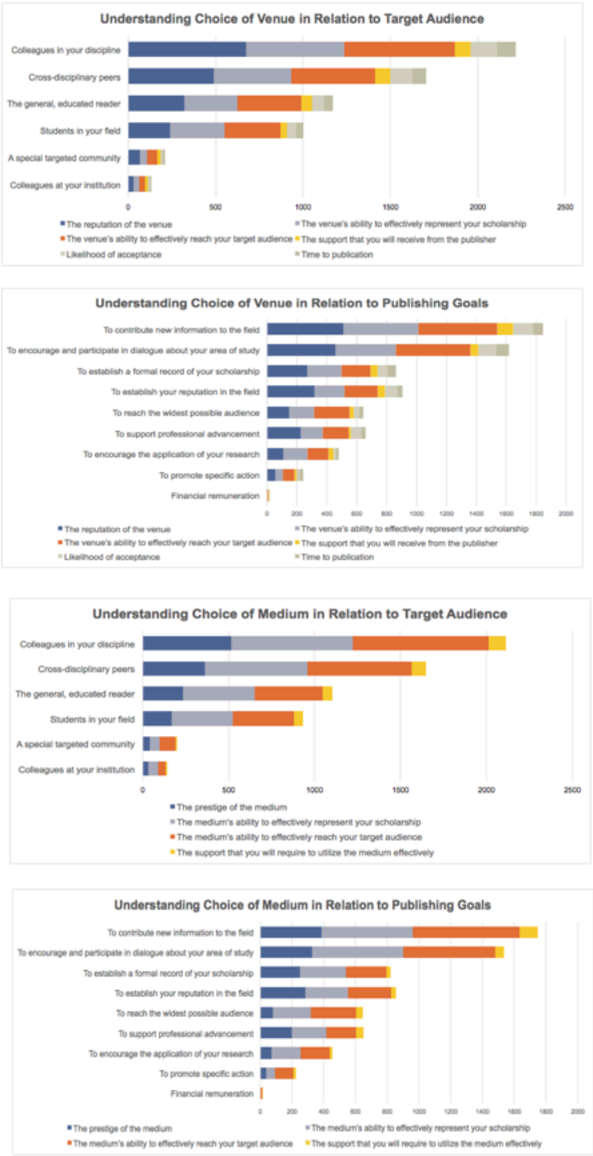


Figure 4. Comparing Scholars Selection Criteria for Publishing Medium and Venue in Relation to Target Audience and Publishing Goals

Discussion

When comparing scholars' characterization of challenges in digital versus print publishing, speed to publication and reaching one's intended audience emerge as the two greatest challenges to print publication, but they are perceived as relatively less challenging in digital formats. This difference between print and digital suggests that these "challenges" might be considered the primary "affordances" that scholars perceive for digital publishing. For digital publishing, the top three challenges that scholars face all relate to receiving adequate support for the logistical aspects of the process, including technical, editorial, and financial support. Though not one of the top three challenges, another aspect of publication that the survey responses suggest is more challenging in digital than print publication is manuscript preparation. Despite the fact that the most prevalent challenges to digital publishing relate to issues of support, the support that scholars will receive from publishers never emerges as a major factor in a scholar's choice of publishing medium or venue, regardless of their specific publication goals and intended audiences.

The top three considerations with respect to choosing both one's medium and the venue in which to publish are the ability to effectively represent the scholarship, the ability to reach one's target audience, and the reputation or prestige of the venue or medium. The weight of these and remaining factors, however, shifts when analyzed in conjunction with scholar's goals and target audiences, suggesting opportunities for developing more nuanced consultative support services when selecting tools and platforms in light of scholars' goals and intended audience. For digital publishing, the first two considerations (representation and audience) are likely to be the determining factor in a scholar's decision to shift away from traditional print publishing and to consider library-based digital publishing opportunities. The emphasis on reputation and prestige, however, may prove problematic for fledgling initiatives that seek to develop alternatives to the established publication models of university presses. Further research will investigate what constitutes acceptable markers of prestige to determine the importance of affiliation with an institution of higher education, which libraries already have, or affiliation with known university presses, which most libraries do not have.

Conclusion

Compared to other publishing models, situating support for scholarly communication in the research library creates possibilities for addressing challenges related to access and sustainability of digital scholarly publishing. This support can be performed efficiently as a part of library activities, leveraging pre-existing technical infrastructure that is designed to support discovery and preservation as well as digital scholarships programs within scholars' commons. These aspects of library-based publishing prove especially compelling in light of survey findings that the biggest challenges for digital publishing include securing adequate technical support services, in addition to financial and editorial support.

The Publishing Without Walls initiative is seeking to offer attractive solutions for authors 1) whose scholarship is not sufficiently represented in the print medium and 2) who place a high value on the technological affordances provided by open access digital scholarship to reach their intended audiences. We further anticipate that developing value-added support services in the form of individual consultations and incubation workshops will help ease the support-related challenges cited by scholars, particularly when assessing which platforms and tools will best represent an author's scholarship.

Bibliography

Courant, P.N. and Jones, E.A. (2015). "Scholarly publishing as an economic public good." In M. Bonn & M. Furlough (eds), *Getting the Word Out: Academic Libraries as Scholarly Publishers*. ACRL. https://www.alastore.ala.org/detail.aspx?ID=11378 (accessed 29 October 2016).

Fenlon, K., Bonn, M., Green, H., Maden, C., Senseney, M., and McCollough, A. (2016). "Understanding the needs of scholars in a contemporary publishing environment." *Proceedings of the American Society for Information Science and Technology*, 53(1).

Lippincott, S. (2016). *Library Publishing Directory 2016.* Library Publishing Coalition. http://www.librarypublishing.org/resources/directory/lpd2016 (accessed 29 October 2016).

Missionaries, Politicians, and Boy Bands: Onlooker Behavior on Twitter During the Nepal, Kumamoto, and Ecuador Earthquakes

David Shepard
shepard.david@gmail.com
UC Los Angeles, United States of America

Takako Hashimoto
takako@cuc.ac.jp
Chiba University, Japan

Hiroshi Okamoto
hiroshi.okamoto299792458@gmail.com
RIKEN Brain Science Institute, Japan

Tetsuji Kuboyama
kuboyama@tk.cc.gakushuin.ac.jp
Gakushuin University, Japan

Kilho Shin
kilhoshin314@gmail.com
Hyogo University, Japan

Introduction

When an earthquake struck Nepal in 2015, the band One Direction sent tweets encouraging their fans to donate to relief efforts, while an Indian activist tweeted accusations of Christian missionaries trading conversions for aid. While Twitter users were quick to bring their own agendas to the Nepal earthquake, does the same hold true for earthquakes in other parts of the world? A series of earthquakes that struck Kumamoto, Japan, and then Muisne, Ecuador in 2016 attracted a substantial amount of Twitter attention as well, yet as far as we are aware, the One Direction fans and the Indian activist made no comment. These users are onlookers to all three earthquakes: in other words, they are not directly affected by these events, but they tweet about them.

This paper explores onlookers' responses across three different earthquakes: the 2015 Nepal earthquake, and the nearly-simultaneous earthquakes in Kumamoto and Ecuador in 2016, which we treat as a single event. This present work expands on our previous conclusion that onlookers tend to bring their own agendas to disasters. This paper shows that users who tweeted about the Kumamoto and Ecuador earthquakes were generally more interested in the earthquake or the affected areas than their own agendas, as their interest in the earthquake could not be predicted by interests in other topics.

Background

A substantial amount of research has explored how social media causes users to engage with political, social, and humanitarian problems; however, opinions on social media's effectiveness—whether it causes users to donate money or participate in campaigns—are mixed. Some argue that displaying concern in social media is more about acquiring social capital than effecting change (Shulman; Gladwell; Morozov, The Net Delusion; Morozov, To Save Everything, Click Here), while a Pew Research Center survey finds that social media does create change (Raine, Purcell, and Smith). One analysis found that charities' use of social media does not increase donations (Malcolm), while another finds that certain tweeting strategies do (Gasso Climent) although tweets may not raise awareness about the charity's causes (Bravo and Hoffman-Goetz). All these studies concur that social media enable substantial discourse about crises. The question we explore here is how much of this conversation is predicted by a user's preexisting interests, and how this varies even among the same type of event in different areas.

Methodology

We followed a similar data collection process for both Nepal and the Kumamoto and Ecuadorean earthquakes: we sampled data from Twitter's REST API to attain a broad sample of onlookers. For Nepal, we had gathered a dataset of tweets sent during the three weeks following the Nepal earthquake by searching for any tweets that mentioned the word "Nepal" from April 24, 2015 to May 8, 2015. We then randomly selected 15,000 users from this set and harvested all tweets they sent between April 24, 2014 and May 8, 2015. We attempted to capture only English-speaking users to increase the likelihood that we would capture users not directly affected by the earthquake, but we still found some users who tweeted in multiple languages. This left roughly 11,000 onlookers for Nepal. For Kumamoto and Ecuador, we gathered a dataset of tweets sent in the two weeks following the Kumamoto earthquake that mentioned "Kumamoto" or "earthquake." We randomly selected 30,000 users and harvested every tweet they sent between March 16 and May 16, 2016. We collected more users, but fewer tweets for each user, than we did in the Nepal dataset so as to look for users who displayed a broader set of interests. This left around 25,000 onlookers in Kumamoto and Ecuador. We were able to filter out non-English tweets much more effectively in the latter dataset than the Nepal dataset.

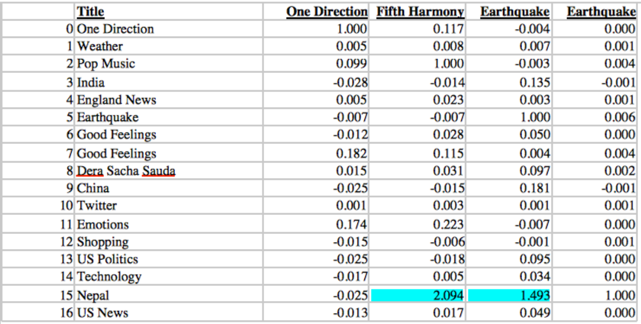
For the tweets for each event, we made a bipartite graph of users to words, and performed community detection using a method proposed by Okamoto and Qiu (2015) [2], which allows for overlapping communities. Okamoto and Qui's method takes a single parameter, alpha, which controls the resolution of community detection: the smaller its magnitude, the larger the number of detected communities. We set alpha to 0.001 in both cases. The output of this method was a list of each node (users and words), and a percentage ranking rating its affinity with each community. We used these results to generate a list of top words in each community, which told us what users who tweeted about that community were interested in. From this process, a number of topics emerged, which we labelled manually according to our interpretations of the top words in each.

Since this method also gave us a ranking for users' affinities to each community, it allowed us to examine the influence of other topics on a user's likelihood to tweet about either event. We wanted to examine how much a user's propensity to tweet about other topics predicted the probability that he or she would tweet about topics related to the earthquake. We ran multivariate linear regressions on each topic in the dataset using the Python sklearn module (Pedregosa et al.). We ran one regression for each topic, in which we treated a user's propensity to tweet about the topic under consideration as a dependent variable predicted by his or her propensity to tweet about other topics.

Results

Our analysis demonstrated a certain predictive power for some topics in each dataset. Applying this process to the Nepal tweets produced 17 topics about a variety of concerns, from entertainment to world events. Table 1 shows these topics. Two of them, topics 5 and 15, treat the earthquake directly.

A correlation exists between tweeting about entertainment topics and tweeting about the earthquake. Tweeting about topic 15 predicts that a user will tweet about topic 2, which is about pop music: the top words include "fifth," "harmony," "video," and "Justin." This correlation is the strongest in the dataset; few other topics show nearly as much correlation. Consequently, we observe a degree of correlation between tweeting about entertainment topics and tweeting about the disaster in Nepal. While the more targeted topics, like the One Direction topic, do not show much correlation with other topics, the more general entertainment topic does.

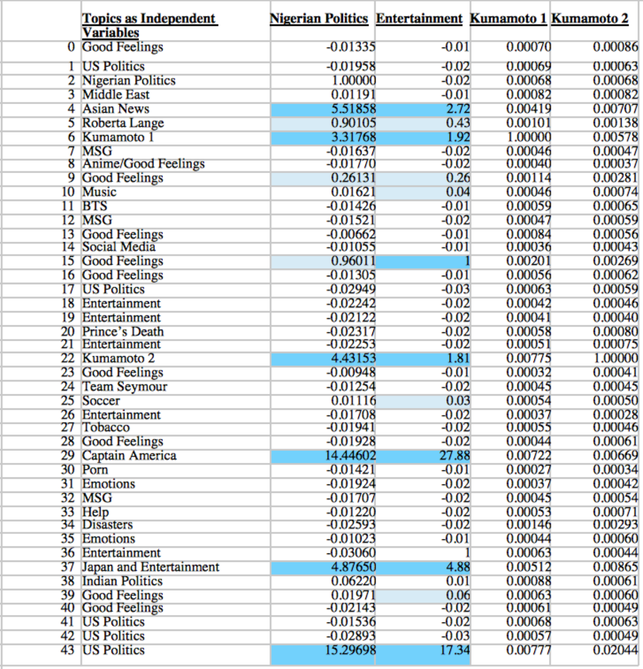


	Title	One Direction	Fifth Harmony	Earthquake	Earthquake
0	One Direction	1.000	0.117	-0.004	0.000
1	Weather	0.005	0.008	0.007	0.001
2	Pop Music	0.099	1.000	-0.003	0.004
3	India	-0.028	-0.014	0.135	-0.001
4	England News	0.005	0.023	0.003	0.001
5	Earthquake	-0.007	-0.007	1.000	0.006
6	Good Feelings	-0.012	0.028	0.050	0.000
7	Good Feelings	0.182	0.115	0.004	0.004
8	Dera Sacha Sauda	0.015	0.031	0.097	0.002
9	China	-0.025	-0.015	0.181	-0.001
10	Twitter	0.001	0.003	0.001	0.001
11	Emotions	0.174	0.223	-0.007	0.000
12	Shopping	-0.015	-0.006	-0.001	0.001
13	US Politics	-0.025	-0.018	0.095	0.000
14	Technology	-0.017	0.005	0.034	0.000
15	Nepal	-0.025	2.094	1.493	1.000
16	US News	-0.013	0.017	0.049	0.000

Table 1: Topics for Nepal, showing probability of tweeting about one topic (X-axis) given likelihood of tweeting about other topics (Y-axis)

In summary, we observe correlation between tweeting about entertainment topics and tweeting about the Nepal earthquake. Those who bring other agendas such as an interest in a particular musical group to the disaster tend to tweet mostly about those topics.

Does the same hold true for Kumamoto and Ecuador? Table 2 shows a few topics from the Kumamoto and Ecuador earthquakes. Our analysis demonstrates that an onlooker's propensity to tweet about some topics could be predicted by interest in others. For example, a user who tweeted about news topics, such as U.S. politics (specifically, topic 43) or Asian news (topic 4), was likely to tweet about Nigerian politics (topic 2). Likewise, a user who tweeted about Japanese Entertainment (37) was also likely to tweet about other entertainment topics.



	Topics as Independent Variables	Nigerian Politics	Entertainment	Kumamoto 1	Kumamoto 2
0	Good Feelings	-0.01335	-0.01	0.00070	0.00086
1	US Politics	-0.01958	-0.02	0.00069	0.00063
2	Nigerian Politics	1.00000	-0.02	0.00068	0.00068
3	Middle East	0.01191	-0.01	0.00082	0.00082
4	Asian News	5.51858	2.72	0.00419	0.00707
5	Roberta Lange	0.90105	0.43	0.00101	0.00138
6	Kumamoto 1	3.31768	1.92	1.00000	0.00578
7	MSG	-0.01637	-0.02	0.00046	0.00047
8	Anime/Good Feelings	-0.01770	-0.02	0.00040	0.00037
9	Good Feelings	0.26131	0.26	0.00114	0.00281
10	Music	0.01621	0.04	0.00046	0.00074
11	BTS	-0.01426	-0.01	0.00059	0.00065
12	MSG	-0.01521	-0.02	0.00047	0.00059
13	Good Feelings	-0.00662	-0.01	0.00084	0.00056
14	Social Media	-0.01055	-0.01	0.00036	0.00043
15	Good Feelings	0.96011	1	0.00201	0.00269
16	Good Feelings	-0.01305	-0.01	0.00056	0.00062
17	US Politics	-0.02949	-0.03	0.00063	0.00059
18	Entertainment	-0.02242	-0.02	0.00042	0.00046
19	Entertainment	-0.02122	-0.02	0.00041	0.00040
20	Prince's Death	-0.02317	-0.02	0.00058	0.00080
21	Entertainment	-0.02253	-0.02	0.00051	0.00075
22	Kumamoto 2	4.43153	1.81	0.00775	1.00000
23	Good Feelings	-0.00948	-0.01	0.00032	0.00041
24	Team Seymour	-0.01254	-0.02	0.00045	0.00045
25	Soccer	0.01116	0.03	0.00054	0.00050
26	Entertainment	-0.01708	-0.02	0.00037	0.00028
27	Tobacco	-0.01941	-0.02	0.00055	0.00046
28	Good Feelings	-0.01928	-0.02	0.00044	0.00061
29	Captain America	14.44602	27.88	0.00722	0.00669
30	Porn	-0.01421	-0.01	0.00027	0.00034
31	Emotions	-0.01924	-0.02	0.00037	0.00042
32	MSG	-0.01707	-0.02	0.00045	0.00054
33	Help	-0.01220	-0.02	0.00053	0.00071
34	Disasters	-0.02593	-0.02	0.00146	0.00293
35	Emotions	-0.01023	-0.01	0.00044	0.00060
36	Entertainment	-0.03060	1	0.00063	0.00044
37	Japan and Entertainment	4.87650	4.88	0.00512	0.00865
38	Indian Politics	0.06220	0.01	0.00088	0.00061
39	Good Feelings	0.01971	0.06	0.00063	0.00060
40	Good Feelings	-0.02143	-0.02	0.00061	0.00049
41	US Politics	-0.01536	-0.02	0.00068	0.00063
42	US Politics	-0.02893	-0.03	0.00057	0.00049
43	US Politics	15.29698	17.34	0.00777	0.02044

Table 2: Topics for Kumamoto and Ecuador, showing probability of tweeting about one topic (X-axis) given likelihood of tweeting about other topics (Y-axis) (truncated for space)

On the other hand, no such correlation was observed in the opposite direction: no topic predicted a user's tendency to tweet about topics 6 and 22, the earthquake topics. All coefficients in those regressions were under 0.01. The two topics that focus on Kumamoto are relatively closed: users who tweet most about the Kumamoto earthquake tweet about little else during this period.

Our interpretation is that users who tweeted about Kumamoto or Ecuador were specifically interested in earthquakes, Japanese culture, or the affected regions. The majority of users who tweeted about the Kumamoto and Ecuador earthquake topics were interested in specialized topics relevant to the events: they were not, for example, One Direction fans. We therefore conclude that while some users tweeting about Kumamoto and Ecuador were motivated by general interests in news or entertainment, they were a much smaller group than in the Nepal dataset.

Conclusions

We find that while users often brought their own agendas to tweeting about Nepal, fewer did so when tweeting about Ecuador and Japan. Users who tweeted about Kumamoto and Ecuador tended to focus on topics related to the earthquakes, and less on issues that the earthquakes might demonstrate.

Our future work will test these conclusions with other earthquakes. In particular, we will examine the 2011 Tohoku Earthquake which raised serious political issues. Additionally, in our present work, we treat the Kumamoto and Ecuador earthquakes as a single event because distinct "Kumamoto" and "Ecuador" topics did not emerge from our text mining, which itself is suggestive of how Twitter users understood them. In our future work, we will probe more deeply for differences between the two earthquakes.

Bibliography

Bravo, C. A., and Hoffman-Goetz, L. (2015) "Tweeting About Prostate and Testicular Cancers: What Are Individuals Saying in Their Discussions About the 2013 Movember Canada Campaign?" *Journal of Cancer Education*. 1–8. link.springer.com. Web. 20 Feb. 2016.

Gasso Climent, C. (2015) "Twitter as a Social Marketing Tool : Modifying Tweeting Behavior in Order to Encourage Donations." info:eu-repo/semantics/bachelorThesis. N.p., 27 Aug. 2015. Web. 20 Feb. 2016.

Gladwell, M. (2011) *Outliers: The Story of Success.* Reprint edition. Back Bay Books. Print.

Malcolm, K. (2016). "How Social Media Affects the Annual Fund Revenues of Nonprofit Organizations." *Walden Dissertations and Doctoral Studies* (2016): n. pag. Web.

Morozov, E. (2012) *The Net Delusion: The Dark Side of Internet Freedom*. Reprint edition. New York: PublicAffairs. Print.

Morozov, E. (2014) *To Save Everything, Click Here: The Folly of Technological Solutionism.* First Trade Paper Edition edition. New York: PublicAffairs. Print.

Okamoto, H., and Qiu, X.-L. (2015) "Modular Decomposition of Markov Chain: Detecting Overlapping and Hierarchically Organized Communities in Networks." *Abstracts of NetSci-X*. Rio de Janeiro, Brasil: N.p. Print.

Pedregosa, F. et al. (2011) “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–2830. Print.

Raine, L., Purcell, K., and Smith, A. (2016) “The Social Side of the Internet | Pew Research Center.” Pew Research Center: *Numbers, Facts and Trends Shaping Your World.* N.p., 1 Mar.Web. 21 Feb. 2016.

Shulman, S. W. (2009) “The Case Against Mass E-Mails: Perverse Incentives and Low Quality Public Participation in U.S. Federal Rulemaking.” *Policy & Internet 1.1*: 23–53. Wiley Online Library. Web. 21 Feb. 2016.

Placing Segregation

Robert Shepard
robert-shepard@uiowa.edu
University of Iowa Libraries, United States of America

Residential segregation and socioeconomic inequality among neighborhoods is common in many American cities. Many large American cities became modern industrial urban places during the mid- to late-nineteenth century, so it is critical to understand the reality of human dimensions in cities during this period. Popular narratives tracing the history of the phenomenon back to that time have been constructed from analyses that used city wards or other large unit areas that aggregate data for many hundreds or even thousands of families. For example, prominent works like *American Apartheid* have concluded that black Americans were not particularly segregated from whites before 1900, and even then they were only slightly more separated from the native white population as the Irish and German immigrants (Massey and Denton 1993). By relying on aggregate datasets from the past, urban historians have been missing out on a level of granularity that would provide better measurements of segregation as well as critical context of spatial separation, given the relative smallness of American cities during their earlier years.

Mid-nineteenth century U.S. Census records contained an abundance of potential information relating to socioeconomic realities of each home during this key formative period in American history. The 1860 and 1870 censuses were especially unique in that they collected each individual's personal estate value and real estate value (these economic variables were omitted from subsequent censuses). However, American census takers did not record address numbers of those they visited before 1880, which was after the Bureau of the Census stopped collecting these critical economic data about individuals and households. Furthermore, even if address data were available on manuscript census records for the time period, much of the built environment has changed since the mid nineteenth century, including the complete renaming of roads and – most importantly – the renumbering of houses along the street grid.

Rather than using easily-available aggregate data collected at city ward levels to make inferences about past urban geographies, this work has combined city directories and period advertisements with census records to reconstruct digitally the historical address systems of cities and geolocate every possible family in the 1860 census for the cities of Washington, D.C., Nashville, Tennessee, and, for the 1870 census, the city of Omaha, Nebraska (additional cities will be added as the project expands). Notably, Donald DeBats and Mark Lethbridge had pioneered a similarly intensive approach for the much smaller cities of Newport, Kentucky and Alexandria, Virginia (DeBats and Lethbridge 2005). Because of the extensive details collected by these censuses, these geolocated individuals provide rich new datasets for historical researchers to explore. By geolocating individuals, this research also highlights specific individual living situations and helps to complement known (and more easily accessible) anecdotal accounts from common primary sources such as journals and letters.

Placing Segregation is the primary result of this research effort - a new, open access digital project at the University of Iowa Libraries' Digital Scholarship and Publishing Studio that explores research questions about housing segregation and socioeconomic disparities across nineteenth century American cities through a series of fully interactive maps and scholarly interpretations derived from the geolocated census data. This presentation introduces core functionality of the digital exhibit and also explains in detail the process of developing the data and the website.

Built primarily upon Leaflet.js, an open-source JavaScript library for mobile friendly interactive maps, as well as other openly available JavaScript libraries such as Fuse.js, this digital map exhibit gives public audiences direct access to detailed census data for the years 1860 and 1870 in these select U.S. cities. A search feature allows visitors to quickly locate the historical residences of persons of interest and read complete census information about them, while a number of simple layer filters permit users to explore potentially infinite topics by giving them direct control over the geographic data. The primary digital exhibit gives audiences without GIS experience the power to ask research questions in a spatial environment.

Initial research findings from the project indicate that urban historians have been substantially underestimating the degree of housing segregation experienced by free black residents, even though some past research built on ward-level analyses closely approximated the extent of housing segregation between the native white population and various immigrant groups. The data generated for this project illustrate American cities were in fact significantly divided according to social class and wealth, long before the rise of the automobile. However, residents were less divided by wealth differences than they were separated spatially by skin color in Washington D.C. or by German and Irish origin in Nashville and Omaha.

Bibliography

DeBats, D. A., and Lethbridge, M. (2005). GIS and the City: Nineteenth-century Residential Patterns. *Historical Geography.* Vol 33: 78-98.

Massey, D., and Denton, N. A. (1993). *American Aparheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press.

Character-distinguishing features in fictional dialogue: the case of *War and Peace*

Daniil Skorinkin
skorinkin.danil@gmail.com
National Research University
Higher School of Economics, Russia

Introduction

The study of character speech is a topic of fairly consistent interest among digital literary scholars. It is usually acknowledged that voices of characters are essentially different from narrator's own voice and should be treated separately. Some researchers have fictional dialogue removed from the texts they studied before any tools of computational investigation are applied (Hoover, 2004). Quite a lot of effort has been made recently to address the problem of identifying character speech in prose and attributing it to the correct speaker (ссылки!). One of the outcomes of such research is the possibility to study voices of different characters on relatively large scale and apply computational tools that measure their recurring stylistic parameters.

Method

The study of character speech has traditionally had strong ties to the fields of stylometry and authorship attribution, as their methods proved quite useful for studying idiolect of a fictional speaker. Suffice it to say that one of the seminal works in stylometry, Computation into criticism# by Burrows (Burrows, 1987), was focused on the study of character speech in Jane Austen's novels. The method developed by Burrows grew into what is currently known as Delta, a widely-adopted standard for authorship attribution. Delta has been consistently and successfully applied to identifying the author of an unattributed text of different languages and genres, but at the same time it saw considerable usage as a purely stylometric tool for the study of text where authorship is undisputed. Among other things, this included research into the specific idiolects of fictional characters (see, for example, Rybicki, 2006).

In our research Delta was used as one of the two possible approaches to studying character voices in Leo Tolstoy's War and peace. Much like in case of Senkewic (Rybicki, 2006), there's certain critical opinion (Eikhenbaum, 2009) that Tolstoy's characters are quite distinct from each other in their speech. Our own experience of carefully reading speech instances extracted from War and peace (for details on extraction procedure see (Skorinkin, Bonch-Osmolovskaya, 2015) supports the opinion. So it seemed natural to try and test computational methods that already showed their applicability to precisely such task. We used R package stylo by (Eder *et al,* 2013)

Testing the method on Russian material

Surprisingly enough, we were unable to find any work that applied Delta to any Russian material. Therefore we felt obliged to conduct a couple of experiments that would test its general applicability to Russian before we proceed with character speech. At the first stage we tried Delta's ability to distinguish between Tolstoy and Dostoevsky. The training set contained one of the six parts of Dostoyevsky's Crime and Punishment and three of the fifteen books of Tolstoy's War and Peace. The remaining 18 pieces of text (5 by Dostoevsky and 13 by Tolstoy) constituted the test set. The results with different settings can be seen in Table 1 and Figures 1,2:

N most frequent	Words	Character 3-grams	Character 3-grams	Character 3-grams
25	80% (4/5)	60% (3/5)	60% (3/5)	100% (5/5)
30	80% (4/5)	80% (4/5)	60% (3/5)	80% (4/5)
35	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
40	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
45	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
50	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
55	80% (4/5)	60% (3/5)	80% (4/5)	80% (4/5)
60	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
65	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
70	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
75	80% (4/5)	80% (4/5)	80% (4/5)	80% (4/5)
80	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
85	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
90	100% (5/5)	80% (4/5)	80% (4/5)	100% (5/5)
95	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
100	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)

Table 1. Delta authorship attribution, Tolstoy vs Dostoevsky

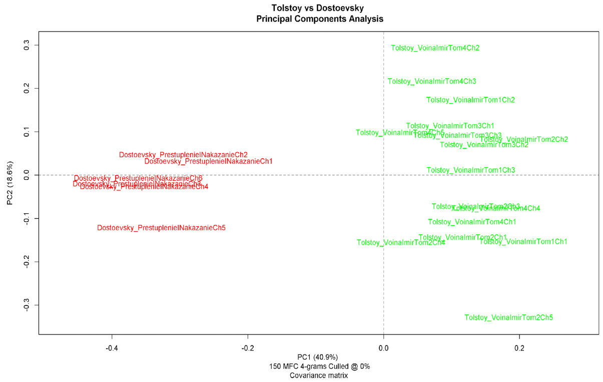


Fig. 1. Delta PCA on 150 most frequent character 4-grams, Tolstoy vs Dostoevsky

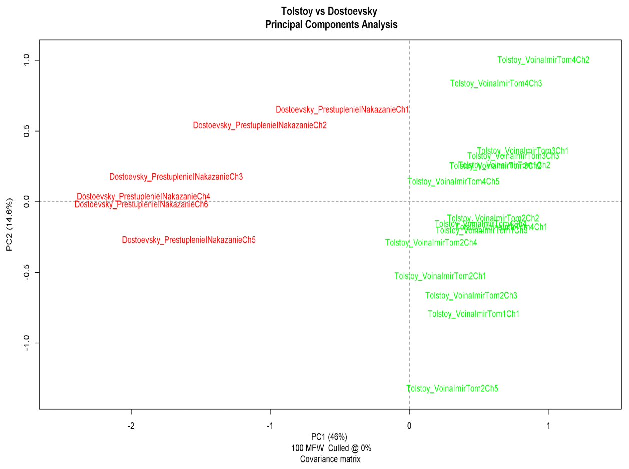


Fig. 2. Delta PCA on 100 most frequent words, Tolstoy vs Dostoevsky

The second experiment involved four Russian authors Tolstoy, Dostoevsky, Goncharov and Turgenev. All four represent (roughly) the same epoch of Russian literature and all four are recognized as masters of realistic prose. We used three novels by each author for our experiment. At the first stage two out of each three were placed in the training corpus, and Delta was supposed to attribute the remaining one. All four novels from the test corpus were attributed correctly. At the second stage we reverted the experiment and left only one novel by each author in the training set. In this case Delta consistently showed 7 out of 8 correct attributions (the only mistake being Tolstoy's Family Happiness incorrectly attributed to Dostoevsky.A possible explanation could be that Family Hap-piness is written in first person from the point of view of a young woman, something uncommon for Tolstoy; and the only Dostoevsky's work the training corpus contained was The Insulted and Humiliated , also a first-person narrative). Fig. 3 shows Delta scores for all the novels visualized with help of principal component analysis.

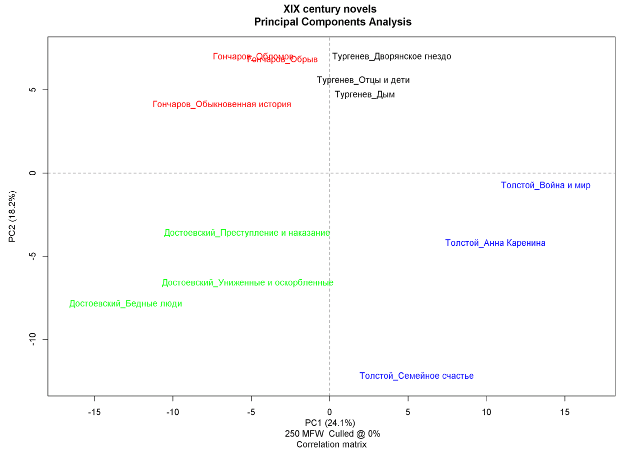


Fig. 3. Delta PCA for 12 Russian novels of 1850-1870-ies, 250 most frequent words

Applying Delta to Tolstoy's characters

Having thus shown that Delta is applicable to Russian, we proceeded with our experiment. In the first place we applied the method to top 5 characters by the total number of speech instances. We split the total sets of speeches by each character and then tried authorship attribution The results are shown in Table 2.

N most frequent	Words	Character 3-grams	Character 3-grams	Character 3-grams
25	80% (4/5)	60% (3/5)	60% (3/5)	100% (5/5)
30	80% (4/5)	80% (4/5)	60% (3/5)	80% (4/5)
35	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
40	80% (4/5)	60% (3/5)	60% (3/5)	80% (4/5)
45	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
50	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
55	80% (4/5)	60% (3/5)	80% (4/5)	80% (4/5)
60	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
65	100% (5/5)	60% (3/5)	80% (4/5)	80% (4/5)
70	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
75	80% (4/5)	80% (4/5)	80% (4/5)	80% (4/5)
80	80% (4/5)	60% (3/5)	80% (4/5)	100% (5/5)
85	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
90	100% (5/5)	80% (4/5)	80% (4/5)	100% (5/5)
95	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)
100	80% (4/5)	80% (4/5)	80% (4/5)	100% (5/5)

Table 2.

The most common mistakes are between princess Marya Bolkonskaya and Natasha Rostova and between prince Andrew Bolkonsky and Pierre Bezukhov. Their closeness can be seen in Figure 4:

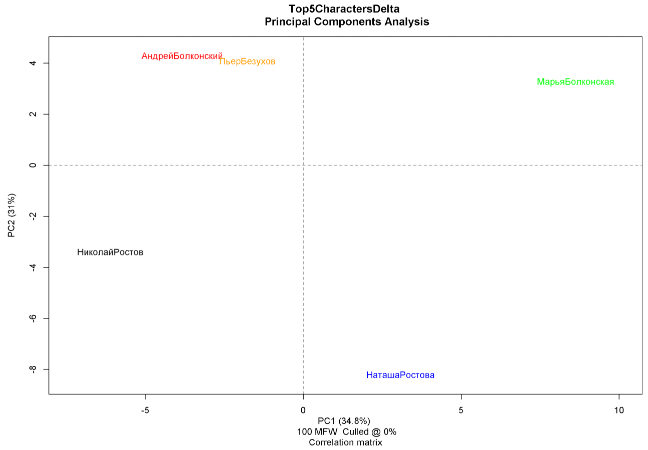


Fig. 4. Delta PCA for top 5 most talkative characters in War and Peace, 100 most frequent words

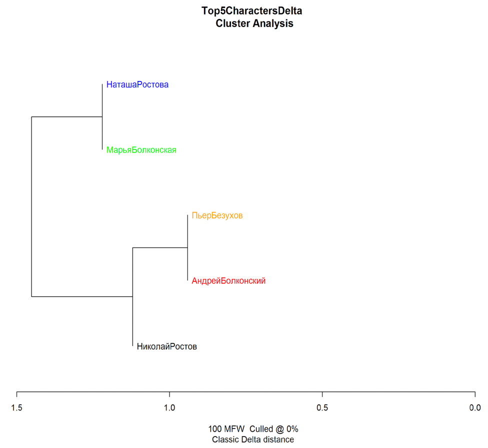


Fig. 5. Delta-based hierarchical clustering for top 5 most talkative characters in War and Peace, 100 most frequent words

The quality of speech authorship attribution inevitably got worse once we expanded our selection from 5 to 15 characters. The results were still quite tolerable reaching 10 out of 15 with certain settings. The analysis of mistakes showed that a) they're less likely to occur between characters of different gender and b) the mistaken characters have quite a lot of mutual conversations.

Further on, we decided to pay more attention to overall Delta scores of character voices and see if they give us any meaningful clustering of characters. Figure 6 shows PCA of characters based on Delta.

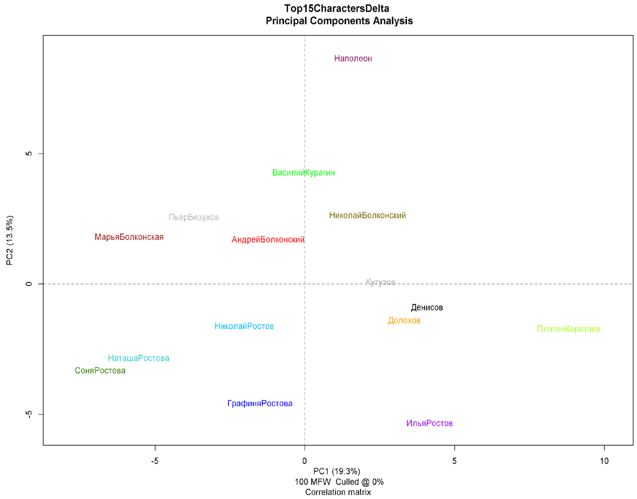


Fig. 6. Delta-based PCA for top 15 most talkative characters in War and Peace, 100 most frequent words

One can easily see the clustering of Rostov family, to a lesser extent this applies to Bolkonsky family as well. Dolokhov, Denisov and Kutuzov could constitute the 'war' cluster.

We then made another expansion and moved from 15 to 30 characters. Figure 7 demonstrates PCA of Delta scores for this selection.

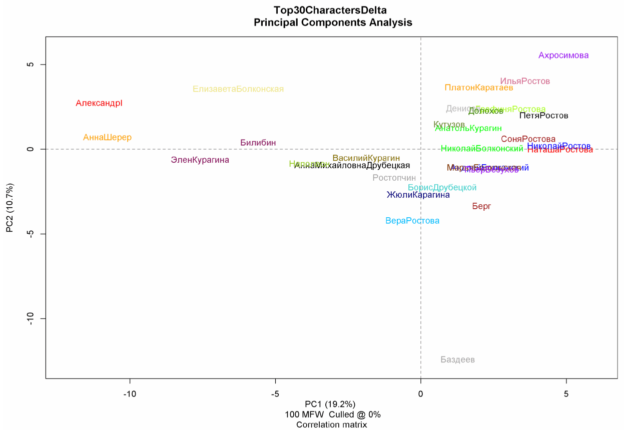


Fig. 7. Delta-based PCA for top 30 most talkative characters in War and Peace, 100 most frequent words

The most striking thing here is the obvious separation of Vera Rostova from the rest of Rostov family. The difference between cold, tempered and rational Vera and her emotional and very sentimental relatives is outspoken and obvious to the reader, but it seems valuable to have this potential quantiative support in the form of different Delta scores. What is even more striking is that Vera is quite close to Berg, a rationalizing careerist who becomes her husband. Note also the closeness of Boris and Julie Karagine, another pragmatic couple happily united in a marriage of convenience.

Applying alternative features

Having tried Delta, we proceeded with our own set of alternative features for character voice analysis (a typical step, as dectibed in Eder, 2015). These features are not related to the lexical makeup of character speech and attempt to reduce the influence of gender-related morphological features of Russian language and the factor of mutual interactions between characters. At this stage we limited ourselves to four features only: the average number of words, the ratio of exclamatory sentences, the ratio oa question sentences and the ratio of punctuation marks (latter being a crude approximation of the 'disruptedness' of speech, which seems rather typical of certain more emotional and lively characters).

When the character set is limited to 5 characters these features even manage to distinguish character speech with some tolerable accuracy (though worse than Delta). However, the analysis of mistakes shows that they capture fundamentally different types of similarities than Delta does. For instance, joyful Natasha in this case is never mistaken for sentimental and melancholic Marya, but rather for her boisterous brother Nikolai. Pierre, on the other hand, is mistaken for Marya rather than for Andrey, who is distinct from them all. Figures 8 and 9 show the results of PCA and hierarchical clustering for these characters based on our own alternative features.

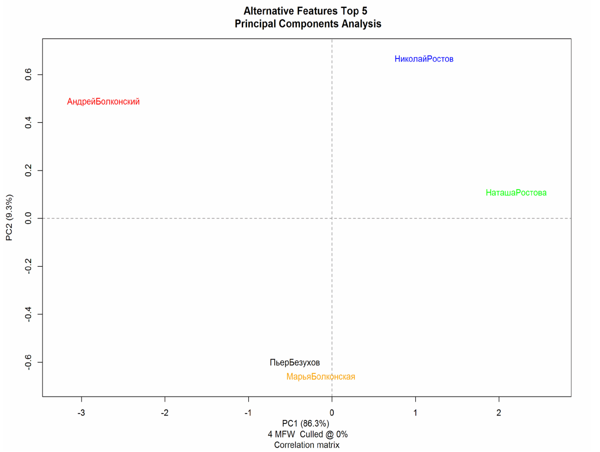


Fig. 8. PCA for top 5 most talkative characters in War and Peace, 4 alternative features

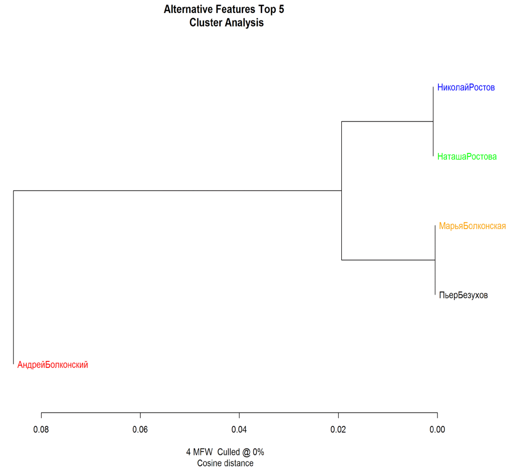


Fig. 9. Hierarchical clustering for top 5 most talkative characters in War and Peace, 4 alternative features

If we compare figures 8 and 9 to their counterparts from the Delta experiment (figures 4 and 5) we can see that the alternative features ignore gender or mutual interactions/ The hypothesis is that they enable us with a more indepth view of a character personality, his/her emotional type and so on.

Figures 10-12 show data on wider selections of characters using alternative features.

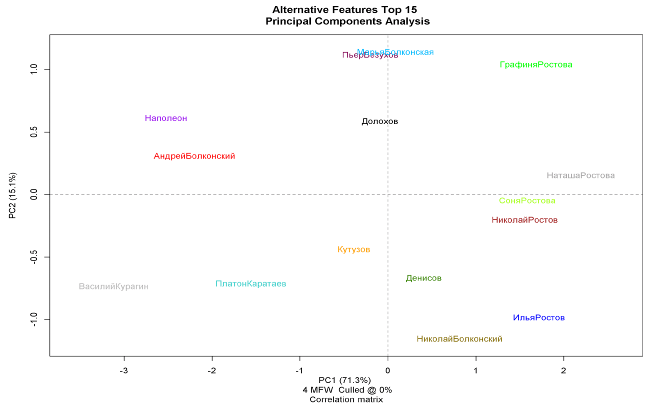


Fig. 10. PCA for top 15 most talkative characters in War and Peace, alternative features

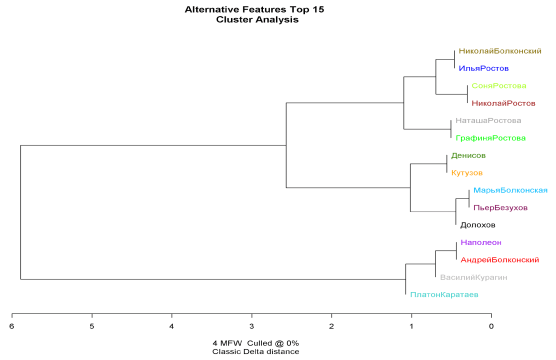
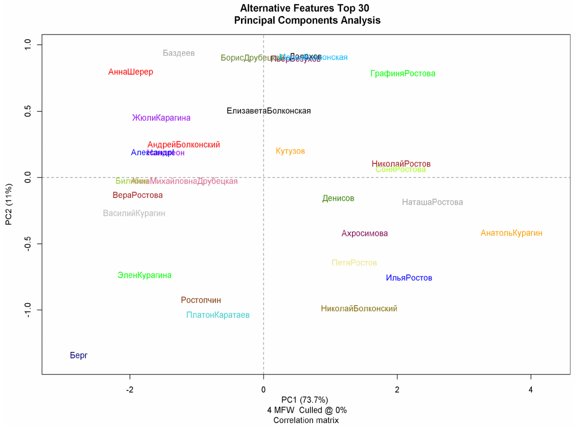


Fig. 11. Hierarchical clustering for top 15 most talkative characters in War and Peace, alternative features

Note that here we do not see any similarity between Andrey and Pierre. Moreover, Andrey is close to Napoleon, which is rather striking given Napoleon is his hero and role model for a considerable part of the novel.



The separation of Vera, on the other hand, is still rather visible - she is far from Moscow-centered Rostov world and close to Saint-Petersburg world of Kuragine family and berg.

Bibliography

Burrows, J. (1987). Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Oxford U. Press.

Eder, M. Kestemont, M. and Rybicki, J. (2013). Stylometry with R: a suite of tools. In: "Digital Humanities 2013: Conference Abstracts". University of Nebraska-Lincoln, Lincoln, NE, pp. 487-89.

Hoover, D.L. (2004) "Testing Burrows's Delta." Literary and Lingusitic Computing 19, no. 4 (2004): 453-475.

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. Literary and Linguistic Computing 21(1), 91-103.

Illustrations to Photographs: Using computer vision to analyse news pictures in Dutch newspapers, 1860–1940

Thomas Smits
t.smits@let.ru.nl
Raboud University, The Netherlands

Most digital humanities projects are based on the analysis of text. However, in our increasingly visually orientated world, it has become clear that we should also devise ways to analyse visual material. In the last couple of years, the Royal Library of the Netherlands (KB) has made important steps in this emergent field. Delpher, the interface that provides access to the KB's digital collections, gives users the opportunity to search for images with captions in its database of digitized newspapers. The KBK-1M database, which holds all the images published in the KB's digitized newspapers, provides researchers with the opportunity to analyse the visual material of this collection in a viable way.

In my paper for DH2017 I will present the preliminary results of my researcher-in-residence project at the KB (six months starting May 2017). My project applies a new computer vision technique to sort the images of the KBK-1M database according to the way in which they were reproduced (engraving/half-tone), thus shedding a new light on an important transitional phase in the history of the visual culture of the news.

The visual representation of news events is generally connected to the technological progress of photography (Bardoel and Wijfjes, 2015). The so-called half-tone revolution of the early 1880s, enabling the massive reproduction of photographs in print media, is seen as forming the basis for our current visual news culture. From a media archaeological perspective, several historians of nineteenth-century media have challenged this technocentric narrative (Gitelman and Pingree, 2003). Hill and Schwartz (2015: 3) propose a contingent history of "news pictures" as a separate "class of images," which does not solely focus on photographic technologies, but on the discourse surrounding them. In relation to this theoretical development, several studies have demonstrated that photography was not the first medium used to visually represent the news. From the early 1840s, illustrated newspapers disseminated news pictures on a massive scale and developed a discourse of objectivity, based on eyewitness accounts, which would be adapted and used for photographs later in the century (Keller, 2013; Gervais, 2010; Barnhust and Nerone, 2000; Park, 1999).

If we accept that the perceived objective nature of news pictures is not based on the affordances of a certain technology but on the discourse(s) of objectivity surrounding a specific medium, it follows that the turning point between the use of illustrations and photographs as the preferred medium to visually represent the news is a critical moment in the history of modern visual news culture. Most commonly, Dutch researchers have presented this point as a watershed, located at the publication of the first photograph of a news event in a newspaper (Kester and Kleppe, 2015). However, case studies from a media archaeological perspective suggest a relatively long transitional period in which illustrations and photographs coexisted and competed as authentic, objective visual representations of the news (Keller, 2013; Steinsieck, 2006). It remains unclear precisely when photography achieved its pre-eminence and why this happened: why did newspapers stop using illustrations to represent the news when they had done so for many decades?

The early reliance on case studies to describe the transitional phase is understandable, as in pre-digital times, a distant reading of the large number of images published in newspapers was all but impossible. My project will shed more light on this important debate by analysing news pictures in Dutch newspapers on a large scale. Using the power of the Dutch supercomputer Cartesius, I will apply the technique of a recent project of Fyfe & Ge (2016) to the images in the KBK-1M database. Fyfe set out to study how computer vision and image processing techniques could be adapted for large-scale interpretation of British Victorian illustrated newspapers (Fyfe, 2016). Using MATLAB, Ge devised a method to analyse two so-called low-level features of images: the pixel ratio, the number of low-intensity pixels divided by the total number of pixels, and the entropy level: the amount of information contained in the image. By juxtaposing these two features, they were able to sort the images of the illustrated newspapers according to the technique used for their reproduction. Half-tones, used to reproduce photographs, exhibit both a high pixel ratio and a high entropy level, while engravings, used to reproduce illustrations, display lower pixel ratios and entropy levels (Ge, 2016). Fyfe has shown that this technique can also be used to identify other categories of images. Maps, for example, exhibit lower pixel ratios and entropy levels than detailed illustrations (Fyfe, 2016).

At DH2017 I will present the first results of my project. By analysing the low-level features of images in the KBK-1M database, I will be able to show when Dutch newspapers started to printed both illustrations and photographs on a large scale. In addition, the period when they competed as objective visual representations of the news can be identified. In doing so, my project introduces a digital humanities approach to the relatively theoretical field of nineteenth-century visual culture studies.

Bibliography

Bardoel, J., and Wijfjes, H. (2015). "Journalistieke cultuur in Nederland. Een professie tussen traditie en toekomst." In J. Bardoel and H. Wijfjes (eds.), *Journalistieke Cultuur in Nederland.* Amsterdam: AUP, pp. 11-29.

Barnhurst, K. and Nerone, J. (2000). "Civic Picturing vs. Realist Photojournalism. The Regime of Illustrated News, 1856-1901." *Design* Issues, 16(1): 59-79.

Fyfe, P. (2016). Illustrated Image Analytics. How Computers See Illustrated Victorian Periodicals [Powerpoint slide]. Retrieved from https://ncsu-las.org/wp-content/uploads/2016/05/wrm-presentation-slides-4-27-16-paul-fyfe.pdf

Ge, Q. (2016) Computer Vision Techniques for Analysis of Illustrations in 19th Century British Newspapers [Powerpoint slide]. Retrieved from https://ncsu-las.org/wp-content/uploads/2016/05/wrm-presentation-slides-4-27-16-qian-ge.pdf

Gervais, T. (2010). "Witness to War: The Uses of Photography in the Illustrated Press, 1855-1904." *Journal of Visual Culture*, 9(3): 370-384. doi:10.1177/1470412910380343

Gitelman, L., and Pingree, G. B. (2003). *New media, 1740-1915*. Cambridge, MA: MIT Press.

Hill, J., and Schwartz, V. (2015). "General Introduction." In J. Hill and V. Schwartz (eds.), *Getting the Picture: The Visual Culture of the News*. London: Bloomsburry Academic, pp. 1-11.

Keller, U. (2013). *The Ultimate Spectacle: A Visual History of the Crimean War*. London: Routledge.

Kester, B., and Kleppe, M. (2015). "Persfotografie in Nederland. Acceptatie, professionalisering en innovatie." In J. Bardoel and H. Wijfjes (eds.), *Journalistieke Cultuur in Nederland*. Amsterdam: AUP, pp. 53-76.

Park, D. (1999). "Picturing the War: Visual Genres in Civil War News." *The Communication Review,* 3(4): 287-321. doi:10.1080/10714429909368588

Steinsieck, A. (2006). "Ein imperialistischer Medienkrieg. Kriegsbenchterstattung im Sudafrikanischen Krieg (1899-1902)." In U. Daniel (ed.), *Augenzeugen. Kriegsberichterstattung vom 18. zum 21. Jahrhundert.* Göttingen: Vandenhoeck & Ruprecht, pp. 87-112.

Linking the Same Ukiyo-e Prints in Different Languages by Exploiting Word Semantic Relationships across Languages

Yuting Song
gr0260ff@ed.ritsumei.ac.jp
Ritsumeikan University, Japan

Taisuke Kimura
is0013hh@ed.ritsumei.ac.jp
Ritsumeikan University, Japan

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Ritsumeikan University, Japan

Akira Maeda
amaeda@is.ritsumei.ac.jp
Ritsumeikan University, Japan

Introduction

Many libraries and museums around the world have been releasing their digital collections and making them accessible online. They provide new opportunities for people to acquire information, but they also pose new challenges for accessing these large quantities of humanities resources. Language barriers are one of the main issues for accessing multiple databases in different languages. In this paper, we propose a method to link Ukiyo-e prints between databases in different languages by exploiting semantic similarity of metadata values across languages, in order to achieve our ultimate research goal that aims to provide multilingual access to multiple and diverse databases. We believe our proposed method could assist users in accessing Ukiyo-e databases regardless of languages.

Ukiyo-e, Japanese traditional woodblock print, is known as one of the popular arts of the Edo period (1603–1868). Many libraries, museums and galleries in Western countries have digitalized Ukiyo-e woodblock prints with the metadata values in different languages. Table 1 shows an example of the same Ukiyo-e prints that are exhibited in multiple databases with metadata values in different languages.

Ukiyo-e prints	Metadata values		Language	Database
	Title	Artist		
	凱風快晴	葛飾北斎	Japanese	Edo-Tokyo Museum
	Gaifū kaisei	Katsushika Hokusai	English (Transliteration)	Library of Congress
	South Wind, Clear Sky	Katsushika Hokusai	English	Metropolitan Museum of Art
	Vent frais par matin clair	Hokusai Katsushika	French	French Photo Agency
	Helder weer en een zuidelijke wind	Katsushika Hokusai	Dutch	Rijksmuseum
	Fuji bei schönem Wetter von Süden gesehen	Katsushika Hokusai	German	Bildarchiv Foto Marburg

Table 1. An example of the same Ukiyo-e prints that are exhibited in multiple databases with metadata values in different languages

For linking the same Ukiyo-e prints between databases in different languages, our previous methods (Batjargal et al., 2014; Kimura et al., 2015; Kimura et al., 2016) utilize the metadata values to calculate the similarity between Ukiyo-e prints, in which the metadata values are translated into the same language by using bilingual dictionaries or online machine translation services.

Resig (2013) has developed an image similarity based Ukiyo-e print search system, which is able to search the same Ukiyo-e prints from multiple databases regardless of languages. However, this method cannot be applied to other humanities resources that have no images, such as texts, audio, video and so on.

In this paper, we use the metadata values to calculate the similarity between Ukiyo-e prints, which is the same as our previous methods (Batjargal et al., 2014; Kimura et al., 2015; Kimura et al., 2016). The difference is that we calculate similarity between metadata values of Ukiyo-e prints in different languages without translating.

Methodology

Our method is based on word embeddings (Mikolov et al., 2013a), which are dense, low-dimensional and real-valued vectors for representing words. By using word embeddings, the words with a similar meaning have closer distances in a vector space, which means the semantic relationships between words can be captured. An example is shown in Fig. 1, in which two words “storm” and “hurricane” that express similar concepts are closer in a vector space (only two dimensions are shown for simplicity). Word embeddings can be learned by using the Word2Vec toolkit, which employs a simple neural network model that can be trained on a large amount of unstructured text data in a short time (billions of words in hours).

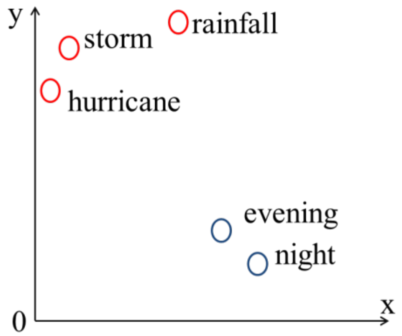


Fig. 1 An example of capturing the sematic relationships between words by using word embeddings

Our proposed method is motivated by the idea of Mikolov et al., (2013b) that the same concepts have similar geometric arrangements across languages. Fig.2 illustrates the vector representations of Japanese words (“雨” and “嵐”) and English words (“rainfall” and “storm”) that are used to describe weather phenomena. It can be seen that the same concepts (e.g. weather phenomena) in Japanese and English have similar geometric arrangements in a vector space.

What is more important is that the relationship between vector spaces that represent these two languages can possibly be captured by learning a mapping between them, e.g. a liner mapping (dotted arrows in Fig.2). If we know some word pairs in Japanese and English, e.g. “雨” and “rainfall”, “嵐” and “storm”, we can learn a mapping that can help us to transform other words in the Japanese vector space to the English vector space.

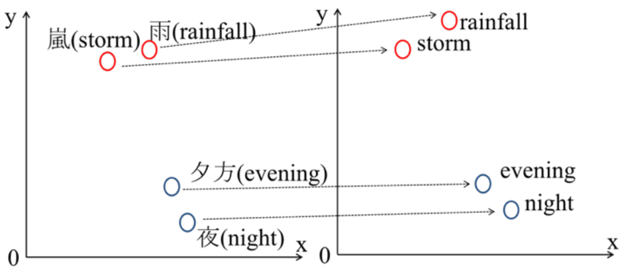


Fig. 2 The vector representations of words that are used to describe weather phenomena (“storm” and “rainfall”) and time (“evening” and “night”) in Japanese (left) and English (right)

Our goal is to measure the similarity between Ukiyo-e prints by using their metadata values in different languages. Motivated by the idea above, we represent textual metadata values as vectors in each language. Then, we learn a mapping between vector spaces that represent two languages in order to transform the vector representations of textual metadata values from source language space to target language space. Once we obtain the vector representations of textual metadata values in target language space, we can calculate the similarity between metadata values in different languages.

Fig. 3 illustrates how our method works. Firstly, we represent the titles of Ukiyo-e prints by additive combination of the vectors of words that compose the titles (Step 1 shown in Fig. 3). And then, we learn the mapping between vector spaces that represent different languages by using some title pairs in Japanese and English (Step 2 shown in Fig. 3), which can help us to transform metadata values from one language space to the other language space.

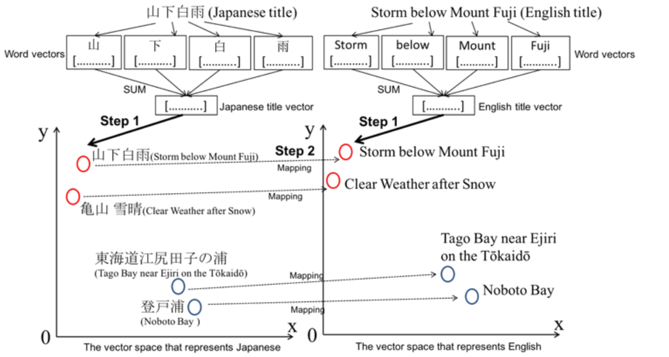

山下白雨 (Japanese title)
Storm below Mount Fuji (English title)
Word vectors
山 [..........]
下 [..........]
白 [..........]
雨 [..........]
Storm [..........]
below [..........]
Mount [..........]
Fuji [..........]
Word vectors
SUM
[..........] Japanese title vector
SUM
[..........] English title vector
Step 1
Step 1
y
y
山下白雨(Storm below Mount Fuji)
Step 2
Mapping
Storm below Mount Fuji
Clear Weather after Snow
Mapping
亀山 雪晴(Clear Weather after Snow)
東海道江尻田子の浦
(Tago Bay near Ejiri on the Tōkaidō)
Mapping
Tago Bay near Ejiri on the Tōkaidō
登戸浦
(Noboto Bay)
Mapping
Noboto Bay
0
x
0
x
The vector space that represents Japanese
The vector space that represents English

Fig. 3 An example that illustrates the main tasks of the proposed method

Experiments

We conducted experiments to evaluate our proposed method in linking the same Ukiyo-e prints in Japanese and English.

In the experiments, the titles are used to calculate similarities among Ukiyo-e prints. Based on our method, the Japanese and English titles are represented by additive combination of the vectors of words that compose the titles. We train the Japanese and English word vectors on

Japanese and English Wikipedia articles using Word2Vec toolkit.

In the process of learning the mapping between two language spaces, we use 600 Japanese-English parallel short sentence pairs for pre-training the mapping between Japanese and English vector spaces. In order to make this mapping more suitable to Ukiyo-e titles, we further use 74 pairs of Japanese and English Ukiyo-e titles to optimize this mapping, in which each pair of titles refers to the same Ukiyo-e prints. This optimized mapping are used to transform other vectors of the titles in Japanese space to English space.

We calculate the similarities between the titles of Ukiyo-e prints using cosine similarity. For each Japanese title, after we obtain the mapped vector in English space, our method outputs the most similar English title vector as its corresponding English title.

We use 173 pairs of Japanese and English Ukiyo-e titles as the test data to evaluate our method. The precision at top-n are used to measure the experimental results, which means the percentage of Japanese titles whose truly corresponding English title are ranked in top n. In order to verify the effectiveness of using Ukiyo-e titles to optimize the mapping, we show the results of both conditions of using Ukiyo-e titles and without using them in the pre-training. The experimental results are shown in Table 2.

	Precision in top-1	Precision within top-5	Precision within top-10	Precision within top-15
Without using Ukiyo-e titles	2.3%	12.2%	17.4%	22.7%
Using Ukiyo-e titles	29.1%	41.9%	50.0%	54.7%

Table 2. The experimental results

These results show that the precisions are further improved by using Japanese and English Ukiyo-e titles to optimize the mapping between Japanese and English vector spaces. The experimental results also confirm the usefulness of our proposed method for linking the same Ukiyo-e prints in Japanese and English.

Conclusion

Our proposed method measures the similarity between metadata values without using any bilingual dictionary or online machine translation system. Moreover, our proposed method represents the metadata values using word embeddings, which can capture the semantic relationships between metadata values.

In the future, we will evaluate our method for linking Ukiyo-e prints in other languages.

Bibliography

Batjargal, B., Kuyama, T., F. Kimura and Maeda, A. (2014). "Identifying the same records across multiple Ukiyo-e image databases using textual data in different languages." *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. London, United Kingdom, pp. 193-96.

Kimura, T., Batjargal, B., Kimura, F. and Maeda, A. (2015). "Finding the Same Artworks from Multiple Databases in Different Languages." *Digital Humanities 2015: Conference Abstracts*. Sydney, Australia.

Kimura, T., Song, Y., Batjargal, B., Kimura, F. and Maeda, A. (2016). "Identifying the Same Ukiyo-e Prints from Databases in Dutch and Japanese." *In Digital Humanities 2016: Conference Abstracts.* Kraków, Poland, pp. 822-24.

Resig, J. (2013). "Aggregating and analyzing digitized Japanese woodblock prints." *In 3rd Annual Conference of the Japanese Association for Digital Humanities.* Kyoto, Japan. https://ukiyo-e.org/about.

Mikolov, T., Chen,K., Corrado, G. and Dean, J. (2013a). "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Le, Q.V. and Sutskever, I. (2013b). "Exploiting similarities among languages for machine translation." *arXiv preprint arXiv:1309.4168*.

Establishing a "Resilient Network" for Digital Humanities

Lisa Spiro
lisamspiro@gmail.com
Rice University, United States of America

Geneva Henry
genevahenry@email.gwu.edu
George Washington University
United States of America

Toniesha Taylor
tltaylor@pvamu.edu
Prairie View A&M University, United States of America

Amanda French
alfrench@vt.edu
George Washington University
United States of America

It can be challenging to be a solo digital humanist, given the range of skills required, deficits in training, and the need for software, hardware and technical support. As E.E. Snyder points out, researchers may be using digital methods but not consider themselves "digital humanists," and thus exist outside of virtual networks like Twitter and conferences like Digital Humanities (2012). Even as colleges and universities recognize the potential of digital humanities, they often struggle to support and sustain DH projects. Digital humanities centers, like many cutting-edge, interdisciplinary academic programs and centers, are fragile, subject to changing priorities and budget cuts (Sample 2010). Moreover, DH centers may not necessarily be able to support the range of researchers' interests and needs. In any case, many digital humanists work at institutions without a DH center, so they often lack institutional support and immediate colleagues whom they can learn from and with. Sample describes the common situation of the DH scholars who lack centers: "We'll never be able to turn to colleagues who routinely navigate grant applications and budget deadlines… We'll never have an institutional advocate on campus who can speak with a single voice to administrators, to students, to donors, to publishers, to communities about the value of the digital humanities" (Sample 2010). The isolation among many digital humanists also means that effort is duplicated, as, for example, faculty at multiple institutions are developing DH educational materials similar to those being created elsewhere.

Networks provide a potential solution to isolation by linking people with shared research interests and enabling them to exchange ideas and expertise. Nancy Maron suggests that a campus-based network model-- whether with equal partners or "a strong central hub, like a library or a DH center, with many spokes" -- may be a preferred organizational model for DH, as it can balance experimentation and sustainability and combine units' strengths (Maron 2015). Beyond the campus, as Sample argues, researchers can build their own communities that transcend institutions and are more agile and resilient than formal organizations: "Stop forming committees and begin creating coalitions. Seek affinities over affiliations, networks over institutes" (2010). Networks can cultivate collective expertise and facilitate acting on common interests without getting caught up in bureaucracy or being limited by long-term obligations.

Indeed, digital humanists are participating in a range of networks, from global to regional to university-based. Many digital humanists are connected around the world through Twitter's virtual network. At the country or regional level, organizations such as Red de Humanidades Digitales (RedHD), NYCDH, and the Texas Digital Humanities Consortium (authors Spiro and Taylor are part of the steering committee of the TDHC) provide online platforms for researchers to discover each other and share information, as well as organize training and events. Within universities or university systems, digital humanities networks such as the Oxford DH Network and the University of Wisconsin-Madison's Digital Humanities Research Network coordinate events and build community among local digital humanists.

As powerful as these networks are, most do not provide funding for collaboration across institutions on research projects, nor do they organize common work on curriculum. Enter Resilient Networks to Support Inclusive Digital Humanities, a collaboration among George Washington University (GW), Rice University, Davidson College, and Prairie View A&M University funded by the Andrew W. Mellon Foundation in the spring of 2016. This network aims to advance digital scholarship by sponsoring collaborative projects among faculty, librarians and students and by developing openly available educational modules that can be used to form a DH curriculum. It brings together private research universities in Washington DC (GW) and Texas (Rice), a public historically black university in Texas (Prairie View A&M), and a private liberal arts college in North Carolina (Davidson College). The co PIs are a member of the library staff and a faculty member from each institution who, together with the project director, oversee the program through bi-weekly online meetings and occasional face-to-face meetings. An Advisory Committee provides strategic guidance for the project.

Between the spring of 2016 and the spring of 2018, Resilient Networks will create:

- A set of openly licensed, adaptable educational modules on digital humanities that can be used in different contexts, such as workshops and semester-long courses. Planned modules include introduction to digital humanities, data in the humanities, the ethos of digital humanities, and framing projects for the public, as well as electives on topics such as text mining and database design and development.
- Cross-institutional projects in which a faculty member, librarian and students collaborate on digital humanities research. To facilitate the cross-institutional projects, Resilient Networks is awarding faculty $5000 jump start packages as seed funding; one is being granted at each institution during year 1, and three during year 2.
- A group of librarians, faculty, and students knowledgeable about digital humanities methods and collaborative approaches. In August 2016, a small group of faculty and librarians gathered at GW for a training workshop on DH project development and humanities approaches to data facilitated by Trevor Munoz. In addition, the Network sponsored a THATCamp at the Digital Frontiers conference hosted at Rice in September 2016 and a THATCamp at George Washington University in March 2017. The network will further support training by organizing THATCamps and providing funding for members to attend intensive DH workshops.
- Intra- and inter-institutional relationships that will facilitate ongoing DH collaborations. Many networks depend on strong personal relationships. Through collaborative work on research, curriculum and training, the Resilient Networks will develop such relationships, laying the foundation for ongoing collaborations.

There are challenges in establishing the Network that are to be expected with a cross-institutional collaboration, including setting common goals, maintaining strong communication, negotiating different academic calendars and bureaucratic systems, and accomplishing tasks in the face of competing responsibilities. In addition, the sheer diversity of digital humanities methods makes it difficult to build a coherent community. On the upside, however, by working within the existing institutional structures at each university rather than creating a separate organizational unit, the work will more likely be sustainable in the long term and better serve the needs of the researchers at each institution. As Snyder cautions, "Decentralised networks that lack both institutional support and dedicated time spent in creating resources will face serious barriers; if there is no position that has explicit responsibility for developing the network, the network may fall by the wayside in the pressure of more urgent responsibilities" (2012). To mitigate this risk, Resilient Networks hired a Digital Humanities Project Director to organize program activities and manage day-to-day operations. Establishing inter-institutional and cross-institutional ties will leverage already-established organizational structures rather than creating new ones. We expect the network to scale to include more institutions, which will expand available expertise. We will be conducting assessments to evaluate the various aspects of the resilient network model to determine its effectiveness in meeting our overall objectives.

In this short paper, we will discuss the network model for digital humanities research and education, results from the first year of "Resilient Networks," and future plans.

Bibliography

Maron, N. (2015). "The Digital Humanities Are Alive and Well and Blooming: Now What?" EDUCAUSE Review, August 17. http://er.educause.edu/articles/2015/8/the-digital-humanities-are-alive-and-well-and-blooming-now-what.

Sample, M. (2010). "On the Death of the Digital Humanities Center." Samplereality. March 26. http://www.samplereality.com/2010/03/26/on-the-death-of-the-digital-humanities-center/.

Snyder, E. E. (2012). "A Framework for Supporting the Digital Humanities: An Alternative to the DH Centre." In Proceedings of the Digital Humanities Congress 2012. *Studies in the Digital Humanities.* Sheffield: HRI Online Publications. https://www.hrionline.ac.uk/openbook/chapter/dhc2012-snyder.

'The Royaltie of Sight': A 3D-GIS recreation of 'prospects' and 'perspectives' within an English designed landscape, c.1550–1660

Elizabeth Eleanor Rose Stewart
lizzie.stewart@uea.ac.uk
Univeristy of East Anglia, United Kingdom

Introduction

In 1624, Henry Wotton highlighted the importance of "the Properties of a well-chosen Prospect", a concept of viewing landscapes, which Wotton dubbed "the Royaltie of Sight" (Wotton, 1624, p.14). Sixteenth- and seventeenth-century English designed landscapes were artificially organised to optimise the 'prospect' whilst reflecting the landowners' individual perspectives. However, little analysis has been attempted into determining the extent and characteristics of 'prospect' and 'perspective' at specific sites. The destruction and modernisation of estate landscapes has hindered their analysis and reshaped perceptions of their appearance and development. In this paper, I will therefore demonstrate how 3D-GIS has the capabilities to change this. For my ongoing PhD research, I am introducing 3D-GIS as a new computational tool for producing 3D representations of estate landscapes but also to subsequently provide the geographical and historical context to analyse the visual experience within them. This methodology will contribute to not only the study of designed landscapes but of historic landscapes generally.

Literature Review

In the study of designed landscapes, country houses, gardens, parks, and the wider estate landscape have been individually and collectively explored and analysed. This is evident from the literature produced in disciplines such as literary history, art history, architectural history, garden history, archaeology, geography, and landscape history. Whilst each approach has its own merits, a noticeable lack of collaboration and acknowledgement of alternative methods has created a "disciplinary vacuum" (Spooner, 2010, p.7). In my research, landscape history is the leading discipline, but it also embraces the contributions of the other disciplines, thus creating an altogether multidisciplinary approach.

Our current understanding is predominantly based on research into well-documented and surviving estates. Consequently, there is a bias against sites which no longer survive or are lacking in evidence. Additionally, they are analysed separately from the wider landscape, which is frequently overlooked despite its significant bearing on the development and utilisation of designed landscapes. Although its contributions have been mentioned, the wider landscape has not been examined in detailed concordance with designed landscapes (Dix, 2011, p.152).

As a result, little research exists into how individual estates were experienced within their landscape context. Whilst the history of the senses, experiences and attitudes towards the landscape have been researched (Rippon, 2012; Hamilakis, 2013; Whyte, 2015), the examination and application of these concepts to historic landscapes, including designed landscapes, is minimal. Therefore, a greater understanding of contemporaries' perceptions and experiences of designed landscapes can be reached by ascertaining how such notions were applied to individual estates.

In this paper, these issues will be addressed and a solution provided. This project utilises computer science as a conduit to address the many disciplines already featuring in the study of designed landscapes. Geographical Information Systems (GIS) is becoming increasingly popular in the spatial humanities, particularly historical and landscape studies (Gregory and Geddes, 2014; Knowles (ed.), 2002), but few have used GIS to analyse individual designed landscapes (Dalton, 2012) and 'prospects' within them (Spooner, 2009). Furthermore, 2D-GIS is typically used, which does not provide a suitable perspective for comprehending the intended experience within these landscapes. On the other hand, Computer Aided Design (CAD) can digitally restore lost landscapes (Virtual Past, n.d.), including designed landscapes (Urmston and Historic England, n.d.), thus providing equal opportunity for their examination alongside surviving specimens. However, the essential difference between GIS and CAD is the handling of spatial data, and CAD is unable to conduct further spatial analysis (Abdul-Rahman and Pilouk, 2007, pp.4–5). Therefore, by incorporating CAD into a GIS environment, 3D-GIS can provide a more immersive perspective for the analyst to explore and analyse these landscapes more appropriately and effectively.

3D-GIS is, therefore, a promising approach for the study of historic landscapes. Regarding designed landscapes, the reconstructive capabilities of 3D-GIS have been demonstrated (de Boer et al., 2011). However, 3D-GIS has only more recently been trialled for the analysis of individual estates. This project builds upon my previous work, looking specifically at the 'prospect' with the use of 'viewshed' analysis (Stewart, 2015). By literally adding a new dimension, 3D-GIS has great potential to improve our understanding of designed landscapes.

Case Study

In this paper, I will present a case study which demonstrates the potential of 3D-GIS. My work adopts a regional approach, focusing on Norfolk, Suffolk and Essex. I will present work from one of these geographical areas. The combination of the aforementioned disciplines will provide the foundations upon which the recreations within CAD and GIS can be supported. This will involve the utilisation of polygons and imported CAD models, based on data extracted from geo-referenced contemporary maps, architectural plans, earthwork plans, fieldwork and archaeological surveys, alongside a range of qualitative historical documents, including estate accounts, inventories and contracts.

Once digitally recreated in 3D-GIS, the estate along with its landscape context can then be subjected to further spatial analysis. For the purposes of this study, the visibility of a 'prospect' can be ascertained using 'viewshed' analysis. To calculate this, topographical data produced by LIDAR is required, with amendments to replicate the historic landscape recreated in 3D-GIS. Viewshed analyses are then conducted from predetermined vantage points. Landowners ensured that 'prospects' could be enjoyed from certain places within their estates, such as the piano nobile of the country house, the rooftop, and from structures and landscape features within the grounds of the estate. Coupled with animations, whereby the act of movement throughout the landscape can be generated, the 'prospects' that contemporaries experienced can be recreated from these positions within the landscape.

Finally, the results can be interpreted in light of the evidence analysed using 'reception theory'; the creativity and reactivity of individuals in response to certain works (McGregor and White, 1990, p.1). This will focus on each individual landowner's possessions, such as published texts or artwork, but also personal correspondence, including letters and diaries. This research will utilise reception theory, which has not yet fully implemented into the study of designed landscapes (Hunt, 2013, p.7), in order to ascertain contemporary 'perspectives' towards these landscapes through the 'prospects' experienced within them. I will present the findings from this case study. This will demonstrate not only the synthesis of the approaches and resources addressed, but will provide the best opportunity to analyse the visual experience and thus improve our understanding of sixteenth- and seventeenth-century designed landscapes.

Conclusion

The potential of 3D-GIS to rekindle the analysis of sixteenth- and seventeenth-century designed landscapes will be evident. The results from this case study, as part of a larger study, will help to explore and examine what these landowners' visual perceptions were. Subsequently, our understanding will improve regarding how individual landscape designs have been influenced by their landowners' thoughts and ideas, and the extent of their individuality or conformity to contemporary fashions in landscape design and appreciation. My project will progress research into designed landscapes whilst demonstrating the methodological benefits of 3D-GIS. As a virtual environment, analytical tool, and database, this approach can subsequently contribute to studies in landscape conservation, heritage management and outreach activities, with scope to benefit research within other areas of the spatial and digital humanities.

Bibliography

Abdul-Rahman, A. and Pilouk, M., (2007), Spatial Data Modelling for 3D GIS, Springer Science & Business Media.

Dalton, C., (2012), Sir John Vanbrugh and the Vitruvian Landscape, Routledge, London.

de Boer, A. et al., (2011), 'Virtual Historical Landscapes' Nijhuis, S., Lammeren, R. V. and Hoeven, F. van der (eds.), Research in Urbanism Series 2(Exploring the Visual Landscape: Advances in Physiognomic Landscape Research in the Netherlands), pp.185–203.

Dix, B., (2011), 'Experiencing the past: the archaeology of some Renaissance gardens', Renaissance Studies 25(1), pp.151–182.

Gregory, I.N. and Geddes, A., (2014), Toward Spatial Humanities: Historical GIS and Spatial History, Indiana University Press, Bloomington.

Hamilakis, Y., (2013), Archaeology and the Senses: Human Experience, Memory, and Affect, Cambridge University Press, New York.

Hunt, J.D., (2013), 'A Reception History of Landscape Architecture' in The Afterlife of Gardens, 2nd ed. Reaktion Books, London, pp.11–32.

Knowles, A.K. (ed.), (2002), Past Time, Past Place: GIS for History, Environmental Systems Research Institute Inc., Redlands, CA.

McGregor, G. and White, R.S., (1990), 'Introduction' in Reception and Response: Hearer Creativity and the Analysis of Spoken and Written Texts, Routledge, London, pp.1–7.

Rippon, S., (2012), Making Sense of an Historic Landscape, Oxford University Press, Oxford.

Spooner, S., (2009), ''A prospect two fields' distance': Rural Landscapes and Urban Mentalities in the Eighteenth Century', Landscapes 10(1), pp.101–122.

Spooner, S., (2010), The Diversity of Designed Landscapes: A Regional Approach c.1660-1830, Unpublished PhD Thesis, University of East Anglia.

Stewart, E., (2015), Recreating Contemporary Views Of Seventeenth-Century Designed Landscapes in Norfolk, Unpublished MA Dissertation, University of East Anglia.

Urmston, P. and Historic England, (n.d.) History of Audley End House and Gardens, http://www.english-heritage.org.uk/visit/places/audley-end-house-and-gardens/history/, [Date Accessed: 27.02.2016].

Virtual Past (n.d.) Virtual Past - Bringing History to Life | Our Expertise, http://www.virtualpast.co.uk/about.php, [Date Accessed: 24.04.2016].

Whyte, N., (2015), 'Senses of Place, Senses of Time: Landscape History from a British Perspective', Landscape Research 40(8), pp.925–938.

Wotton, H., (1624), The Elements of Architecture, London.

Ouvrir les boîtes noires : un outil pédagogique pour une approche critique de la recherche d'information en ligne

Cyrille Suire
cyrille.suire@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle, France

Axel Jean-Caurant
axel.jean-caurant@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle, France

Charles Illouz
charles.illouz@univ-lr.fr
Centre de Recherche en Histoire Internationale et Atlantique (CRHIA), Université de La Rochelle
France

Introduction

Les rapports entre les technologies du numérique et les Sciences Humaines et Sociales (SHS) sont aujourd'hui profondément renouvelés par le développement des Humanités numériques. Les mutations en cours permettent en particulier de repenser les pratiques d'enseignement et d'éducation au numérique. Voir en particulier le très récent ouvrage rédigé de manière collaborative lors du Edcamp qui s'est tenu à Paris les 1er et 2 septembre 2016 (Bourgatte et al., 2016). Parmi ce vaste chantier, la problématique de la recherche et de l'accès à l'information semble primordiale. Elle représente en effet une part importante du travail quotidien des chercheurs et des étudiants en SHS et se trouve profondément bouleversée par le tournant numérique. Celui-ci se matérialise par l'explosion du nombre de ressources disponibles en ligne et par une plus grande hétérogénéité des documents et des moyens d'accès. Si la disponibilité immédiate de documents jadis inaccessibles, faute de moyens ou de temps, est un atout précieux pour la recherche actuelle en SHS, il n'en reste pas moins que les activités de recherche et d'accès à l'information requièrent des compétences pointues et une expérience significative pour être véritablement maîtrisées. Le numérique pose à cet égard des problèmes spécifiques qui doivent être pris en compte dans la formation des étudiants.

Des travaux récents ont montré que la numérisation des documents et leur mise à disposition en ligne étaient loin de permettre un accès universel et transparent au matériau de la recherche (Milligan, 2013). Des problématiques techniques, méthodologiques et cognitives limitent notre compréhension de l'accès à l'information (Cardon, 2013). Sur le plan technique, la distinction généralement opérée entre « information seeking », l'ensemble des processus et pratiques des utilisateurs pour répondre à un besoin d'information et « information retrieval », les méthodes et techniques informatiques qui permettent au système de répondre à ces besoins est fortement significative (Buchanan et al., 2005). Il existe un fossé entre les besoins des utilisateurs et les technologies utilisées pour y répondre. Il est ainsi souvent difficile de comprendre les relations de causes à effets entre les critères de recherche (inputs) saisis par l'utilisateur, et les résultats (outputs) fournis par le système. Ce fossé sémantique se décompose en de multiples problématiques. Sans faire ici une liste exhaustive, les exemples ne manquent pas. Les technologies de reconnaissance optique de caractères (OCR), par exemple, génèrent des erreurs qu'il est complexe d'identifier et de mesurer. L'indexation des documents, quant à elle, repose sur des catégories développées par d'autres, dont les utilisateurs n'ont bien souvent pas connaissance. Les paramètres des algorithmes de pertinence, de personnalisation et de classement des résultats sont inaccessibles alors même qu'ils régissent la manière dont sont générés et classés les résultats des recherches en ligne.

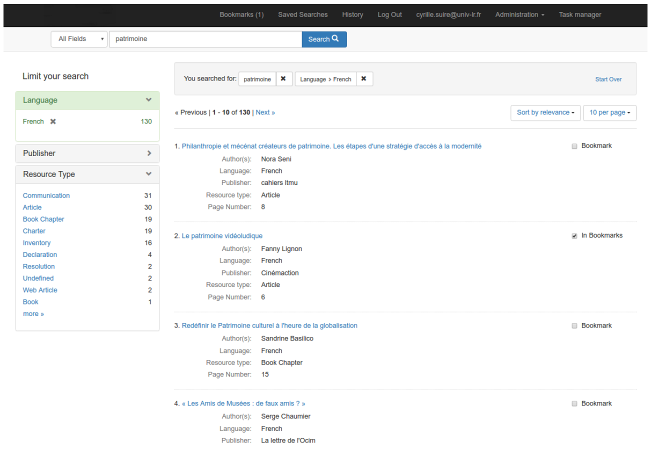
Ces transformations subies par les données (des inputs aux outputs), inconnues des utilisateurs, sont des boîtes noires. Les étudiants doivent avoir conscience de leurs effets et doivent pouvoir les intégrer à leur appareillage critique. L'outil que nous développons poursuit deux objectifs:

Dresser des ponts entre les opérations menées par les systèmes d'accès à l'information et l'idée que s'en font les utilisateurs. Notre outil a ainsi pour vocation de mettre en lumière et permettre d'expliquer l'ensemble des transformations que subissent les données et les documents dans le cadre de l'interaction entre un utilisateur et un système de recherche d'information.

Mettre les étudiants en situation expérimentale de recherche d'information et permettre d'en visualiser les résultats individuels et collectifs. Ces résultats doivent ensuite pouvoir être réinvestis dans une discussion interrogeant la démarche et la méthode de recherche d'information dans un contexte numérique.

Environnement technologique et fonctionnalités

L'outil que nous développons est un moteur de recherche et une interface de bibliothèque numérique reposant sur le framework libre Hydra (Apache Solr, Fedora Commons et Blacklight), capable d'indexer des sources primaires comme secondaires. Il est conçu pour être utilisé lors de séances de cours dirigées par un enseignant où les étudiants doivent répondre à un besoin d'information défini par le formateur.



patrimoine

You searched for: patrimoine ✖ Language › French ✖

Language

French ✖ 130

Publisher

Resource Type

Communication	31
Article	30
Book Chapter	19
Charter	19
Inventory	16
Declaration	4
Resolution	2
Undefined	2
Web Article	2
Book	1

more »

« Previous | 1 - 10 of 130 | Next »

1. Philanthropie et mécénat créateurs de patrimoine. Les étapes d'une stratégie d'accès à la modernité ☐ Bookmark
Author(s): Nora Seni
Language: French
Publisher: cahiers Itmu
Resource type: Article
Page Number: 8

2. Le patrimoine vidéoludique ☑ In Bookmarks
Author(s): Fanny Lignon
Language: French
Publisher: Cinémaction
Resource type: Article
Page Number: 6

3. Redéfinir le Patrimoine culturel à l'heure de la globalisation ☐ Bookmark
Author(s): Sandrine Basilico
Language: French
Resource type: Book Chapter
Page Number: 15

4. « Les Amis de Musées : de faux amis ? » ☐ Bookmark
Author(s): Serge Chaumier
Language: French
Publisher: La lettre de l'Ocim

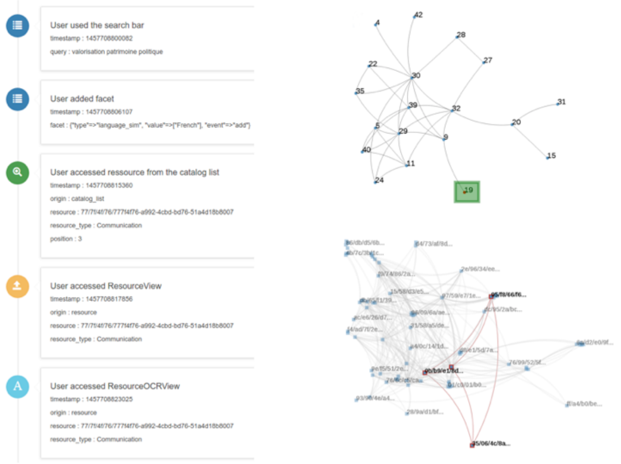
Figure 1 : Interface principale de l'outil de recherche

Sur la base du framework Hydra, nous ajoutons des composants qui documentent la requête effectuée par l'utilisateur et fournissent des informations supplémentaires dont :le résumé de la requête transmise au moteur de recherche (mots, expressions et filtres); les paramètres du moteur (métadonnées interrogées ou ignorées, variations linguistiques utilisées ou ignorées, etc.); les procédures de classement des résultats utilisées (critères de pertinence, pondérations utilisées, etc.).

Cette documentation supplémentaire permet d'afficher à l'utilisateur des compléments d'information pour chaque requête qu'il effectue dans le moteur de recherche. Il lui est alors possible de comprendre comment a été décomposée sa requête, quels paramètres extérieurs à son contrôle ont été appliqués et comment ont été calculés les résultats.

L'enseignant qui dirige une séance peut par ailleurs agir sur certains paramètres du moteur de recherche, pour modifier dynamiquement la manière dont celui-ci réagit aux inputs des utilisateurs. Il est par exemple possible de modifier le comportement de l'algorithme de pertinence, de changer la présentation des résultats ou encore d'activer et désactiver l'usage de certaines métadonnées. Ces options ont un impact important sur le comportement du moteur. Les étudiants peuvent immédiatement le mesurer et réfléchir à l'influence de ces paramètres cachés sur leur pratique de recherche.

Durant la séance, l'outil garde également une trace de l'activité globale de recherche de chaque utilisateur. Cette trace est élaborée grâce aux logs de l'application, que nous enrichissons de données liées au comportement de recherche des utilisateurs (Suire et al., 2016). Ces informations servent d'abord à générer des représentations personnelles et collectives des recherches effectuées durant la séance. Il est ainsi possible de débattre avec les étudiants des résultats obtenus, en se fondant sur les différentes approches et stratégies de recherche qu'ils ont développées. Par ailleurs, les données d'usage collectées par l'application sont utiles à l'évaluation de l'outil. Nous pouvons par exemple comparer le comportement des étudiants avant et après la formation, sur des tâches de recherche d'information de même nature et ainsi évaluer la pertinence pédagogique de nos développements. Nous complétons par ailleurs cette évaluation par des questionnaires et des entretiens qualitatifs, permettant de mesurer l'intérêt des étudiants pour la problématique, l'approche et les outils mis en place.



User used the search bar
timestamp : 1457708800082
query : valorisation patrimoine politique

User added facet
timestamp : 1457708806107
facet : {"type"=>"language_sim", "value"=>["French"], "event"=>"add"}

User accessed ressource from the catalog list
timestamp : 1457708815360
origin : catalog_list
resource : 77/7f/4f/76/777f4f76-a992-4cbd-bd76-51a4d18b8007
resource_type : Communication
position : 3

User accessed ResourceView
timestamp : 1457708817856
origin : resource
resource : 77/7f/4f/76/777f4f76-a992-4cbd-bd76-51a4d18b8007
resource_type : Communication

User accessed ResourceOCRView
timestamp : 1457708823025
origin : resource
resource : 77/7f/4f/76/777f4f76-a992-4cbd-bd76-51a4d18b8007
resource_type : Communication

Figure 2 : Exemples de représentation visuelle: à gauche, un extrait du processus de recherche d'un utilisateur (ici identifié par le numéro 19) et à droite une représentation en réseau des documents consultés par cet utilisateur au regard de l'activité de l'ensemble du groupe.

Contexte expérimental et applications

Bien que l'outil soit encore en développement, nous menons des expériences afin d'évaluer son intérêt pédagogique et son fonctionnement. Nous expérimentons l'outil avec des groupes de 50 étudiants, en 2e année de cycle universitaire (Ces étudiants peuvent être considérés comme débutants, aussi bien en terme de connaissance du domaine qu'en terme de compétences en recherche d'information (Jenkins et al., 2003)), qui débutent leur spécialisation en Histoire. Lors d'une séance de 2 heures, ils bénéficient d'abord d'une courte présentation du fonctionnement de notre outil. Les étudiants ont ensuite 60 minutes pour élaborer une problématique de recherche à l'aide d'un corpus d'environs 300 documents hétérogènes (textes scientifiques ou institutionnels, documents iconographiques, etc.) relatifs à la thématique de la préservation du patrimoine. La séance se termine sur un échange de 40 minutes autour des indicateurs fournis par notre outil.

Lors de nos premières expériences, que nous présenterons plus en détail lors de notre communication, les thématiques de ces échanges ont été nombreuses. A titre d'exemple, les étudiants ont été surpris de leur tendance à se focaliser sur les premiers résultats calculés par le moteur de recherche, alors même que de nombreux documents pertinents se trouvaient dans les pages suivantes. Les capacités de représentations graphiques de notre outil ont également permis de mener une discussion constructive sur les différentes stratégies qu'il convient d'adopter face à ce type de besoin d'information. Les développements en cours permettront prochainement de simuler d'autres situations de recherche d'information courantes en SHS (Ellis, 1989; Savolainen, 2016). Ces premières expériences ont toutefois déjà témoigné de l'intérêt de mettre les étudiants en situation expérimentale, pour les engager dans une pensée critique de la recherche d'information dans un contexte numérique.

Bibliographie

Bourgatte, M., Ferloni, M. and Tessier, L. (2016). Quelles humanités numériques pour l'éducation ? - Éditions MkF http://www.editionsmkf.com/produit/edcamp-icp.

Buchanan, G., Cunningham, S. J., Blandford, A., Rimmer, J. and Warwick, C. (2005). "Information seeking by humanities scholars." International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg, pp. 218-29.

Cardon, D. (2013). Présentation, Réseaux, 177, pp. 9-21.

Ellis, D. (1989). A behavioural approach to information retrieval system design. Journal of Documentation, 45(3): 171-212.

Jenkins, C., Corritore, C. L. and Wiedenbeck, S. (2003). Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. IT & Society, 1(3): 64-89.

Milligan, I. (2013). Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010. The Canadian Historical Review, 94(4): 540-69.

Savolainen, R. (2016). Contributions to conceptual growth: The elaboration of Ellis's model for information-seeking behavior. Journal of the Association for Information Science and Technology, 68(3) : 594-608.

Suire, C., Jean-Caurant, A., Courboulay, V., Burie, J.-C. and Estraillier, P. (2016). "User Activity Characterization in a Cultural Heritage Digital Library System." Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. (JCDL '16). New York, NY, USA: ACM, pp. 257-58.

A Survey of Digital Humanities Programs

Chris Alen Sula
csula@pratt.edu
Pratt Institute, United States of America

S. E. Hackney
s.hackney@pitt.edu
University of Pittsburgh, United States of America

Phillip Cunningham
philc17@gmail.com
The Amistad Research Center, United States of America

Introduction

Existing studies of anglophone digital humanities (DH) curricula have examined course syllabi (Terras 2006; Spiro 2011) and the development of programs at specific institutions (Rockwell 1999; Siemens 2001; Sinclair 2001; Unsworth 2001; Unsworth and Butler 2001; Drucker, Unsworth, and Laue 2002; Sinclair & Gouglas 2002; McCarty 2012; Smith 2014). This study adds to the literature on teaching and learning by presenting a survey of formal degree and certificate programs in anglophone DH.

While these programs represent only part of the entire DH curricula, they are important in several respects: First, they reflect intentional groupings of courses, concepts, skills, methods, and techniques, purporting to represent the field in its broadest strokes. Second, these programs include explicit learning outcomes, and their requirements form one picture of what all DHers are expected to know upon graduation. Third, formal programs organize teaching, research, and professional development in the field; they are channels through which material and symbolic capital flow and thereby shape the field itself. Finally, these programs, their requirements, and coursework are one way—perhaps the primary way—in which prospective students encounter the field and make choices about whether to enroll in a DH program and, if so, which one.

In addition to helping define the field, a study of DH programs also has the capacity to comment on pedagogical discussions in the literature. Hockey, for example, has long wondered whether programming should be taught in the field (1986) and asks, "How far can the need for analytical and critical thinking in the humanities be reconciled with the practical orientation of much work in humanities computing?" (2001). Also skeptical of mere technological skills, Mahony and Pierazzo (2002) argue for teaching methodologies or "ways of thinking" in DH, and Clement examines multiliteracies in DH (e.g., critical thinking, commitment, community, and play), which help to push the field beyond "training" to a more humanistic pursuit (2012, 372). Others have called on DH to engage more fully in critical reflection, especially in relation to technology and the role of the humanities in higher education (Brier 2012, Liu 2012, Walzer 2012).

These and other concerns point to longstanding questions about the proper balance of skills and reflection in DH. While a study of DH programs cannot address the value of critical reflection, it can report on its presence (or absence). These findings, together with our more general observations about DH activities, give pause to consider what is represented in, emphasized by, and omitted from the field at its most explicit levels of educational training.

Methodology

We compiled a list of 37 DH programs active in 2015, drawn from public listings (UCLA Center for Digital Humanities 2015; Clement 2015), background literature, and web searches (e.g., "digital humanities masters"). In addition to degrees and certificates, we included minors and concentrations in which humanities content was the central focus, and omitted digital arts and media programs in which this was not the case. Because our sources and searches are all English-language, it limits what we can say about global DH.

We recorded the URL and basic information (e.g., title, level, location) about each program and looked up descriptions of any required courses in the institution's course catalog. To analyze topics addressed in these programs, we applied the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH 2014), which attempts to capture the "scholarly primitives" of the field (Perkins et al. 2014). TaDiRAH contains forty activities terms organized into eight parent terms ('Capture', 'Creation', 'Enrichment', 'Analysis', 'Interpretation', 'Storage', 'Dissemination', and 'Meta-Activities'). TaDiRAH was chosen for its basis in the literature on "scholarly primitives" (Unsworth 2000), as well as three earlier sources (an arts-humanities.net taxonomy, DIRT categories and tags, and a Zotero bibliography) and community feedback and revision.

We applied terms to program/course descriptions independently and then tested intercoder agreement, which was extremely low. We attribute this to the many terms in TaDiRAH, complexity of program/course description language, questions of scope (i.e., using a broader or narrower term), and general vagueness. We did find discussing our codings helpful and, in doing so, were able to agree. Accordingly, each of us read and coded every program/course description and discussed them until we reached consensus. Often, this involved pointing to specific language in the descriptions and referencing TaDiRAH definitions or notes from previous meetings when interpretations were discussed.

Findings and discussion

The number of DH programs has risen sharply over time, beginning in 1991 and growing steadily by several programs each year since 2008 (see Figure 1).

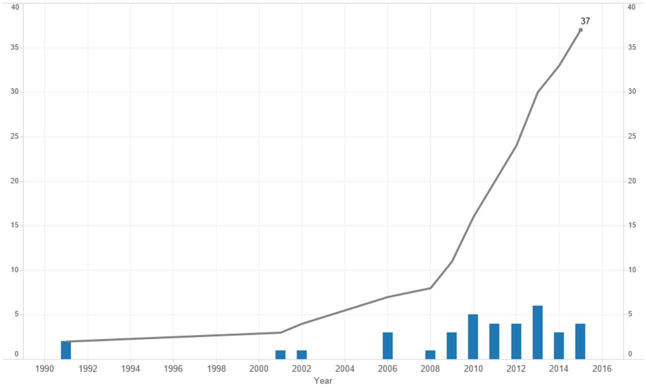


Figure 1. Growth of digital humanities programs in this study

Geography

Most of the programs studied here were located in the US (22 programs, 60%), followed by Canada (6 programs, 16%), the UK (5 programs, 14%), Ireland (3 programs, 8%), and Australia (1 program, 3%). We note that these programs are all located in Anglophone countries and that TaDiRAH, too, originates from this context, which necessarily limits what we can say about DH programs from a global perspective.

Structure

Less than half of these DH programs grant degrees: some at the level of bachelor's (8%), most at the level of master's (22%), and some at the doctoral level (8%) (Figure 2). The majority of these programs are certificates, minors, or specializations—certificates being more common at the graduate level and nearly one-third of all programs studied here.

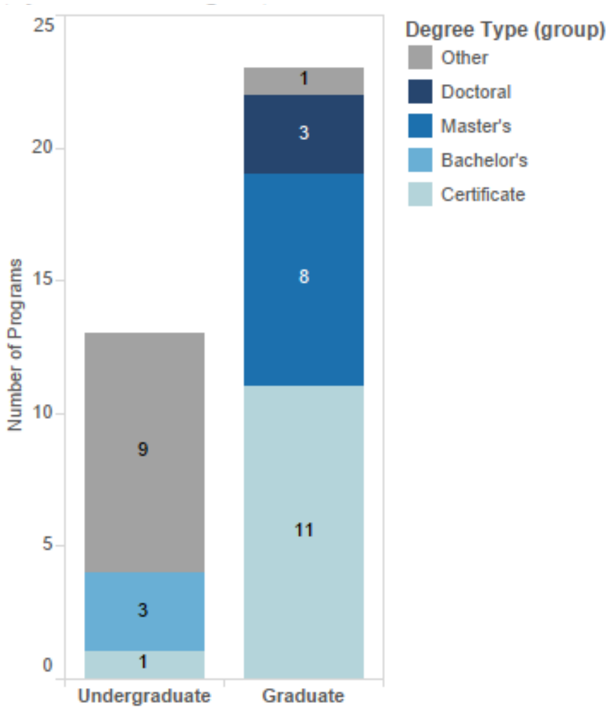


Figure 2. Digital humanities programs in this study (by degree and level)

We also examined special requirements of these programs, finding that half require some form of independent research (see Figure 3), and half require a final deliverable, referred to variously as a capstone, dissertation, portfolio, or thesis (see Figure 4). About one-quarter of these programs require fieldwork, often an internship (see Figure 5).

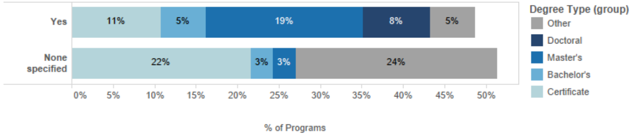


Figure 3. Independent research requirements of digital humanities programs in this study

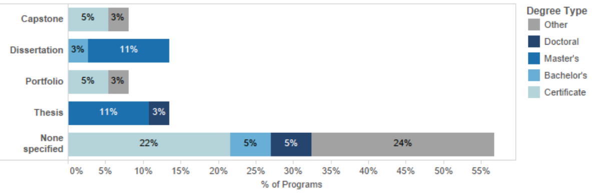


Figure 4. Final deliverable required by digital humanities programs in this study

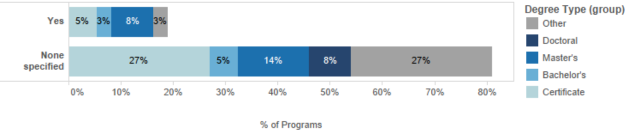


Figure 5. Fieldwork requirements of digital humanities programs in this study

Location & disciplinarity

Most of these programs are housed in colleges/schools of arts and humanities, but about one-third are outside of traditional schools/departments, in centers, initiatives, and, in one case, jointly with the library (see Figure 6). Most DH concentrations and specializations are located within English departments, commensurate with Kirschenbaum's claim that DH's "professional apparatus…is probably more rooted in English than any other departmental home" (2010, 55).

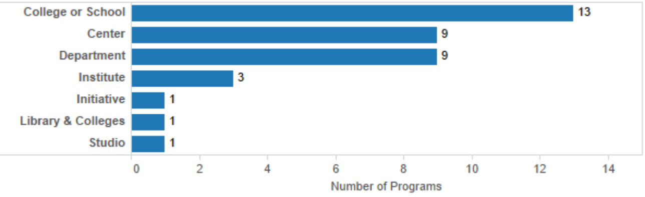


Figure 6. Institutional location of digital humanities programs in this study

Elective courses for these programs span myriad departments and disciplines, from humanities departments (art history, classics, history, philosophy, religion, and various languages) to along with computer science, education, information and library science, design, media, and technology.

DH activities

We analyzed our TaDiRAH codings in two ways: overall term frequency (see Figure 7) and weighted frequency across programs (see Figure 8). To compute weighted frequencies, each of the eight parent terms were given a weight of 1, which was divided equally among the subterms in each area. These subterm weights were summed to show how much of an area is represented, regardless of its size.

Analysis and meta-activities (e.g., 'Community building', 'Project management', 'Teaching/Learning') make up the largest share of activities, along with creation (e.g., 'Designing', 'Programming', 'Writing'). It is worth noting that 'Writing' is one of the most frequent terms (11 programs), but this activity certainly occurs elsewhere and is probably undercounted because it was not explicitly mentioned in program descriptions. The same may be true for other activities.

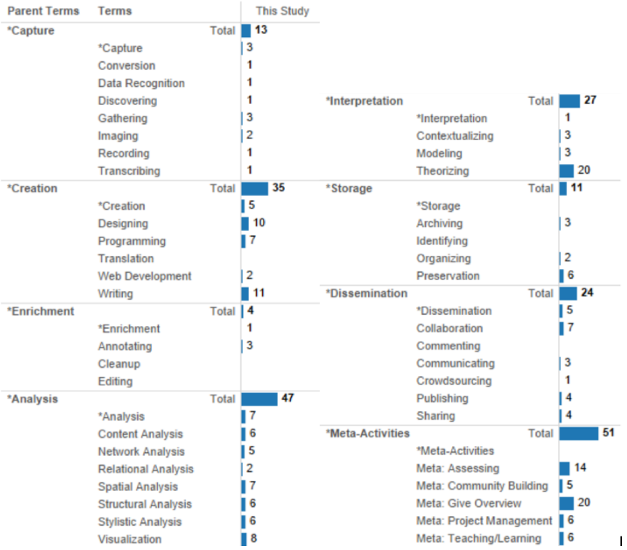


Figure 7. TaDiRAH term coding frequency (grouped)

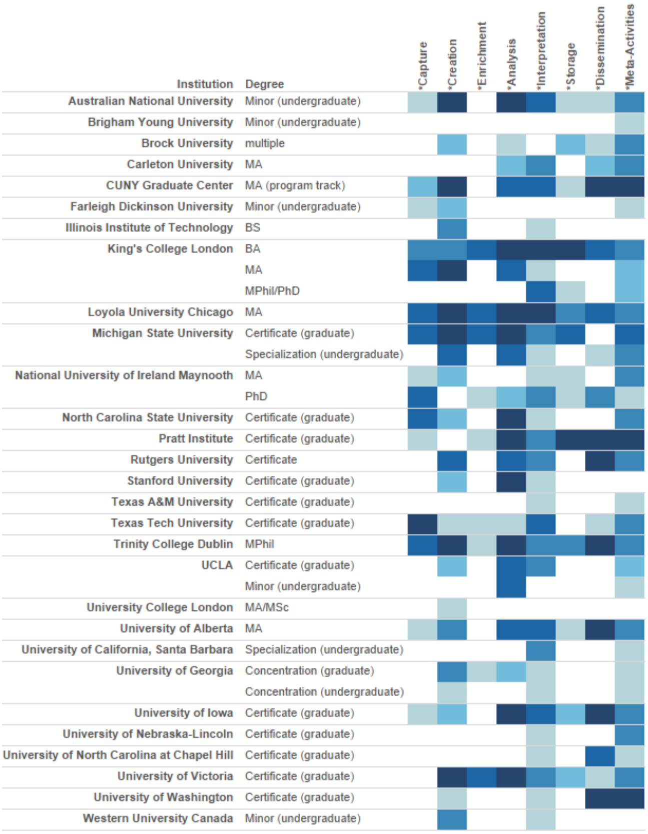
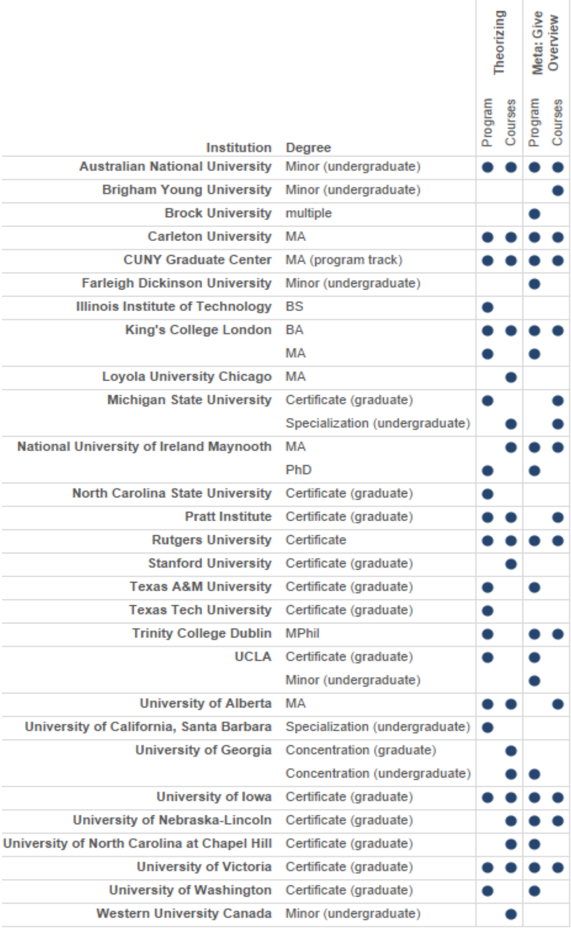


Figure 8. Digital humanities programs in this study and their required courses (by area)

Enrichment and storage terms (e.g., 'Archiving', 'Organizing', 'Preservation') were generally sparse (only 1.9% of all codings), but we suspect these activities do occur in DH programs and courses—in fact, they are assumed in broader activities such as thematic research collections, content management systems, and even dissemination. Generally, there seems to be less emphasis on content ('Capture', 'Enrichment', and 'Storage' terms) and more focus on platforms and tools ('Analysis' and 'Meta-Activities' terms) in the programs studied here—or at least, those may be more marketable to prospective students and institutions, two important audiences of program webpages.

Theory and critical reflection

To analyze theory and critical reflection, we focused our analysis on two terms: 'Theorizing' and 'Meta: GiveOverview', which we used to code theoretical or historical introductions to DH itself. We found that all programs studied here included some mention of theory or historical/systematic overview (see Figure 9). Our codings, of course, do not reveal anything further about the character of this reflection, whether it is the type of critical reflection called for in the literature, or how it interfaces with skills and techniques in these programs.



		Theorizing		Meta: Give Overview	
Institution	Degree	Program	Courses	Program	Courses
Australian National University	Minor (undergraduate)	●	●	●	●
Brigham Young University	Minor (undergraduate)				●
Brock University	multiple			●	
Carleton University	MA	●	●	●	●
CUNY Graduate Center	MA (program track)	●	●	●	●
Farleigh Dickinson University	Minor (undergraduate)			●	
Illinois Institute of Technology	BS	●			
King's College London	BA	●	●	●	●
	MA	●		●	
Loyola University Chicago	MA		●		
Michigan State University	Certificate (graduate)	●			●
	Specialization (undergraduate)		●		●
National University of Ireland Maynooth	MA		●	●	●
	PhD	●		●	
North Carolina State University	Certificate (graduate)	●			
Pratt Institute	Certificate (graduate)	●	●		●
Rutgers University	Certificate	●	●	●	●
Stanford University	Certificate (graduate)		●		
Texas A&M University	Certificate (graduate)	●		●	
Texas Tech University	Certificate (graduate)	●			
Trinity College Dublin	MPhil	●		●	●
UCLA	Certificate (graduate)	●		●	
	Minor (undergraduate)			●	
University of Alberta	MA	●	●		●
University of California, Santa Barbara	Specialization (undergraduate)	●			
University of Georgia	Concentration (graduate)		●		
	Concentration (undergraduate)		●	●	
University of Iowa	Certificate (graduate)	●	●	●	●
University of Nebraska-Lincoln	Certificate (graduate)		●	●	●
University of North Carolina at Chapel Hill	Certificate (graduate)		●	●	
University of Victoria	Certificate (graduate)	●	●	●	●
University of Washington	Certificate (graduate)	●		●	
Western University Canada	Minor (undergraduate)		●		

Figure 9. Theory in digital humanities programs in this study

Further directions

We plan to publish our data and visualizations publicly for researchers, students, and those developing curriculum: http://bit.ly/DHprograms. We believe it provides a baseline of field growth, areas, structure, and learning experiences, which can be used to measure changes in future, in addition to providing a data-driven perspective on the field today.

In that respect, we hope this study gives the community pause to consider how DH is described, represented, and taught. If there are common expectations not reflected here, perhaps we could be more explicit about those, at least in building our taxonomies and describing our formal programs and required courses. Conversely, if there are activities that seem overrepresented here, we might consider why those activities are prized in the field (and which are not) and whether this is the picture of DH we wish to present publicly.

Bibliography

Brier, S. (2012). "Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 390–401. Minneapolis: Univ Of Minnesota Press.

Clement, T. (2012). "Multiliteracies in the Undergraduate Digital Humanities Curriculum." In *Digital Humanities Pedagogy: Practices, Principles and Politics*, edited by Brett D. Hirsch, 365–88. Open Book Publishers. http://www.openbookpublishers.com/product/161/digital-humanities-pedagogy--practices--principles-and-politics.

Clement, T. (2015). "Digital Humanities Inflected Undergraduate Programs." *Tanyaclement.org*. January 8, 2015. http://tanyaclement.org/2009/11/04/digital-humanities-inflected-undergraduate-programs-2/.

Drucker, J., Unsworth, J., and Laue, A. (2002). "Final Report for Digital Humanities Curriculum Seminar." Media Studies Program, College of Arts and Science: University of Virginia. http://www.iath.virginia.edu/hcs/dhcs/.

Hockey, S. (1986). "Workshop on Teaching Computers and Humanities Courses." *Literary & Linguistic Computing* 1 (4): 228–29.

Hockey, S. (2001). "Towards a Curriculum for Humanities Computing: Theoretical Goals and Practical Outcomes." The Humanities Computing Curriculum / The Computing Curriculum in the Arts and Humanities Conference. Malaspina University College, Nanaimo, British Columbia.

Kirschenbaum, M. G. (2010). "What Is Digital Humanities and What's It Doing in English Departments?" *ADE Bulletin* 150: 55–61.

Liu, A. 2012. "Where Is Cultural Criticism in the Digital Humanities?" In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 490–509. Minneapolis, Minn.: Univ Of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/20.

Mahony, S., and Pierazzo, E. (2012). "Teaching Skills or Teaching Methodology." In *Digital Humanities Pedagogy: Practices, Principles and Politics*, edited by Brett D. Hirsch, 215–25. Open Book Publishers. http://www.openbookpublishers.com/product/161/digital-humanities-pedagogy--practices--principles-and-politics.

McCarty, W. (2012). "The PhD in Digital Humanities." In *Digital Humanities Pedagogy: Practices, Principles and Politics*, edited by Brett D. Hirsch. Open Book Publishers. http://www.openbookpublishers.com/product/161/digital-humanities-pedagogy--practices--principles-and-politics.

Perkins, J., Dombrowski, Q., Borek, L., and Schöch, C. (2014). "Project Report: Building Bridges to the Future of a Distributed Network: From DiRT Categories to TaDiRAH, a Methods Taxonomy for Digital Humanities." In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2014*, 181–83. Austin, Texas.

Rockwell, G. (1999). "Is Humanities Computing and Academic Discipline?" presented at An Interdisciplinary Seminar Series, Institute for Advanced Technology in the Humanities, University of Virginia, November 12.

Siemens, R. (2001). "The Humanities Computing Curriculum / The Computing Curriculum in the Arts and Humanities: Presenters and Presentation Abstracts." November 9–10, 2001. https://web.archive.org/web/20051220181036/http://web.mala.bc.ca/siemensr/HCCurriculum/abstracts.htm#Hockey.

Sinclair, S. (2001). "Report from the Humanities Computing Curriculum Conference," Humanist Discussion Group. November 16, 2001. http://dhhumanist.org/Archives/Virginia/v15/0351.html.

Sinclair, S., and Gouglas, S. W. 2002. "Theory into Practice A Case Study of the Humanities Computing Master of Arts Programme at the University of Alberta." *Arts and Humanities in Higher Education* 1 (2): 167–83. doi:10.1177/1474022202001002004.

Smith, D. (2014). "Advocating for a Digital Humanities Curriculum: Design and Implementation." Presented at Digital Humanities 2014. Lausanne, Switzerland. http://dharchive.org/paper/DH2014/Paper-665.xml.

Spiro, L. (2011). "Knowing and Doing: Understanding the Digital Humanities Curriculum." Presented at Digital Humanities 2011. Stanford University.

TaDiRAH. (2014). "TaDiRAH - Taxonomy of Digital Research Activities in the Humanities." *GitHub*. May 13, 2014. https://github.com/dhtaxonomy/TaDiRAH.

Terras, M. (2006). "Disciplined: Using Educational Studies to Analyse 'Humanities Computing.'" *Literary and Linguistic Computing* 21 (2): 229–46. doi:10.1093/llc/fql022.

UCLA Center for Digital Humanities. (2015). "Digital Humanities Programs and Organizations." January 8, 2015. https://web.archive.org/web/20150108203540/http://www.cdh.ucla.edu/resources/us-dh-academic-programs.html.

Unsworth, J. 2000. "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" Presented at Symposium on Humanities Computing: Formal Methods, Experimental Practice, King's College London. http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

Unsworth, J., (2001). "A Masters Degree in Digital Humanities at the University of Virginia." Presented at 2001 Congress of the Social Sciences and Humanities. Université Laval, Québec, Canada. http://www3.isrl. illinois.edu/~unsworth/laval.html.

Unsworth, J., and Butler, T. (2001). "A Masters Degree in Digital Humanities at the University of Virginia." Presnted at ACH-ALLC 2001, New York University, June 13–16, 2001.

Waltzer, L. (2012). "Digital Humanities and the 'Ugly Stepchildren' of American Higher Education." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 335–49. Minneapolis: Univ Of Minnesota Press.

Sharing Surgical Scars: Social Networks and the Many Gendered Meanings of Mastectomy

K. J. Surkan
ksurkan@gmail.com
Massachusetts Institute of Technology
United States of America

In most Western cultural contexts, the surgical modification of the female breast is usually presumed to take place in the context of cosmetic surgery primarily aimed at breast enhancement for the cisgendered female. The cultural obsession with sexualizing breasts as a signifier of both heterosexuality and femininity has meant gender trouble for breast cancer patients (and BRCA "previvors") seeking flat mastectomies, as well as transmasculine, genderqueer, and FTM individuals electing top surgeries. Whereas trans studies critiques have traced the emergence of (and subsequent resistance to) a prescriptive medical narrative authorizing gender affirming top surgeries, only recently have some in the breast cancer community begun to form identity groups around non-reconstructive surgical choices and an embrace of "flatness" on social media. This presentation is a comparative study of the reported experiences and rhetorical choices of social networks of two flat mastectomy groups: those pursuing mastectomy for reasons of gender affirmation, and those for whom mastectomy is primarily a treatment for cancer or cancer risk. The emergence of the Flattopper Pride movement within the breast cancer community represents a queering of conventional (female) breast reconstruction, offering feminist rhetoric that resists presumptions about the "natural" embodiment of femininity while occupying a position that is at once informed by transgender surgical narratives (and counter-narratives) yet articulates a distinctly different relationship to surgical modification of the female body. The circulation of mastectomy images through trans, queer fashion (Cat Walk & Play Out underwear models), and e-patient breast cancer communities represents three distinct yet overlapping discursive contexts in which mastectomy derives cultural meaning. Through an exploration of how mastectomy images are circulated, censored, and shared in a variety of social media networks and support groups, feminist and queer critiques of surgical mastectomy options can be seen to emerge and create cross-influence between breast cancer, transgender, and genderqueer communities.

Bibliography

Tempesta, E. (2015) "Two breast cancer survivors who underwent double mastectomies walk the runway topless during an LGBT fashion show to raise awareness for their fellow 'flattoppers.'" 2 July. *DailyMail.com* Retrieved October 31, 2016. http://www.dailymail.co.uk/femail/article-3147606/Two-breast-cancer-survivors-underwent-double-mastectomies-walk-runway-topless-LGBT-fashion-raise-awareness-fellow-flattoppers.html

Williams, F. (2012) *Breasts: A Natural and Unnatural History.* New York: Norton.

Access(ed) Poetry. The *Graph Poem Project* and the Place of Poetry in Digital Humanities

Chris Tanasescu (MARGENTO)
margento.official@gmail.com
University of Ottawa, Canada

Diana Inkpen
diana.inkpen@uottawa.ca
University of Ottawa, Canada

Vaibhav Kesarwani
vaibhavkw84@gmail.com
University of Ottawa, Canada

Bryan Paget
bdjpaget@gmail.com
University of Ottawa, Canada

The *Graph Poem Project* at University of Ottawa develops tools for poetry computational analysis and applies graph theory and network graph computational apps in structuring, analyzing, and visualizing poetic corpora. The concept involves generating network graphs (multigraphs) in which the vertices are poems and the edges represent various commonalities between poems in terms of subject, diction, form, style, and other criteria. By computationally analyzing the network graph certain extremely interesting and potentially useful information can be extracted regarding a specific corpus. For instance, by identifying cut vertices, we find out which poem(s) play a crucial role in the connectivity of the network and therefore also in the cohesion of the corpus, since by removing that particular poem-node the whole network becomes disconnected, and thus the corpus per se without that particular poem would become disarticulate and divided; and similarly, cut edges signal connections between certain poems that are of paramount importance in the connectivity of the entire network. Identifying cliques of nodes, on the other hand, for instance, shows how poems within a corpus are clustered and which parts of a specific oeuvre, collection, or anthology are more self-contained than others or which poems belong together across or within divides such as authorship, school, period or region, etc. More generally, representing corpora as network multigraphs makes possible analyzing and visualizing both small and significantly large datasets (so far up to hundreds of thousands) of poems, and reach certain information and conclusions about both particular poems in that corpus and the corpus as a whole or as compared to other corpora.

This is one of the first respects in which the issue of access becomes of dramatic importance for our research and for developing and testing our tools; and virtually for any contemporary digital project in poetry. Namely the access to databases and the ability to convert existing files into formats that are compatible to computational processing and analysis. The paper will focus on the existing databases, will review previous work in the field and analyze the kind of data they have employed, will examine the premises and prospects for big data and/or data intensive research in poetry (one of the main tenets of our project), and the arguable absence of such approaches in the published research in the area so far. Terms like “crawl” and “rip” for the poetry available online, and issues related to digitization and/or computationally analyzing poetry in print will also be placed under close scrutiny while we will also report on our own solutions and results, and compare them to those in other projects in digital humanities (DH), digital literary studies, or computational linguistics/analysis.

Our research, tools, publications, and future work will therefore be then presented in a comparative manner in the wider context of current trends, theories, and debates in DH in general and text analysis in particular—by considering for instance the potential relevance of the *Graph Poem Project* to at least some of the issues raised in the “Forum: Text Analysis at Scale” section of *Debates in the Digital Humanities 2016* (eds. Lauren F. Klein and Matthew K. Gold), or literature-related tools/projects such as Syuzhet or heureCLÉA—, in natural language processing (NLP)—by reviewing other projects that have dealt with literary computational analysis and particular poetry processing, but also other works that focused on various issues in computational linguistics, such as syntax or trope processing and analysis—, and in digital pedagogy, by relating it to other poetry related digital pedagogical projects (such as those reviewed by Chuck Rybak, for instance, in *Digital Pedagogy in the Humanities*). “Access” therefore will turn out to encompass quite a few different meanings and phenomena. In terms of text and data analysis, we will review existing debates and commentaries and contribute our own on the politics of developing poetry archives or assembling collections/anthologies, and the related issues of representationality, especially with regards to gender, race, minority, and so forth. In other words, the question asked will be, what is the access of certain authors and their works to our research and tools, what are their chances to be represented in the graph poem? Then, in terms of existing tools, accessibility refers to either the monetary aspects of acquiring or upgrading certain apps or to accessing the source code and/or repurposing it, and the discussion will gravitate around the political and cultural implications of such options. Access then also means accessibility of our tools to the user, in the sense of how easy or how complicated it is for ‘anybody’ to read our approach, use our apps, and employ our results; and why would they ever do it?—how are we trying to ‘seduce’ our audience? The paper will hence examine what are the “compromises” we have made in that respect—as for instance in picking 2 or 3 dimensions as rhyme sub-types in employing a certain API for displaying rhyme scores as scatter graphs—, and what are the actual efforts we are making in conveying the potential significance of sometimes arcane mathematical theories and algorithms for poetry. But on yet another level, access means success, namely the measure of our project’s access to levels of meaning (in Adrian Liu’s terms and not only) that may be deemed as ‘deeply’ or ‘typically’ poetic, an analysis that will combine notions of genre or craft related subtlety with pragmatic computational concerns, such as operationally consolidating numeric and non-numeric features and commonalities.

A more prosaic question to address will then be what is the access poetry really has to DH-related research, interest, and funding. Again, the data related to our own project will be compared to others, and the analysis will focus on the place of poetry in significant DH publications, periodicals, and platforms.

A cross-language comparison of co-word networks in Digital Library and Museum of Buddhist Studies

Muh-Chyun Tang
muhchyun.tang@gmail.com
National Taiwan University, Taiwan

Kuang-hua Chen
khchen@ntu.edu.tw
National Taiwan University, Taiwan

Introduction

This paper reports a co-words domain analysis of Buddhism literature collected by DLMBS (Digital Library and Museum of Buddhist studies) at National Taiwan University. Established in 1995, the DLMBS is one of the most comprehensive online repository of Buddhist research materials. It currently contains over 400 thousand records of books, research papers, theses and dissertations in 45 languages and digitized Buddhist scriptures. A controlled vocabulary, which is in five languages, including

Chinese, English, Japanese, German, and French, was used to help users search DLMBS's bibliographic database. By using co-occurrence data of author assigned keywords in the bibliographic records, this study attempts to generate co-word networks in three different languages, Chinese, English, and Japanese, to compare regional focuses on Buddhist studies.

Co-words analysis has been shown to be effective in mapping the intellectual structure of disciplines (He, 1999; Leydesdroff, 1989). While it has been widely used in the domains of sciences and technologies (e.g. Buitelaar, Bordea, & Coughlan, 2014; Ding, Chowdhury, & Foo, 2001; Bhattacharya & Basu,1998; Looze, & Lemarie, 1997; Peters & van Raan, 1993a, 1993b; Courtial, 1994; Tijssen, 1992; Callon, Courtial, & Laville, 1991; Rip & Courtial, 1984), to the best our knowledge, it has so far not been applied to humanities. Part of the advantage of using co-word in sciences and technologies is the highly codified subject languages, therefore a higher degree of consistency between concepts and terms in these fields. We believe that a cross-language co-word analysis of Buddhist studies literature would be a worthwhile endeavor for a couple of reasons: Firstly, it has been pointed out that there has been a wide variety of methods, perspectives and subject matters within the international communities of Buddhist studies. The heterogeneity of its scholarships can be partly traced to their geographic roots (Cabezón, 1995). It is therefore interesting to empirically study whether and how the intellectual structures reflected in the published literatures in these language communities differ from one another. A comparison of the intellectual structures can shed light on knowledge interests shared and distinct in these three language communities. From the methodological plain, co-word analysis also provides a viable alternative to citation-based network analysis in humanities where the citation structure is known to be much sparser than in sciences and technologies.

Procedures and analysis

Three separate co-word networks were generated in three different languages where nodes denote the keywords and edges the strength of their co-occurrence. Unlike in most of the previous co-words analysis, where keywords were extracted from titles and abstracts, author assigned keywords were used here as it is believed that they are more representative to the content of the articles and tend to have higher degree of consistency than keywords in free-text. For monographs and other types of publications, the subject-headings assigned by human indexers were used as keywords. Edge weights were normalized by both the inclusion and the Jaccard index (Courtial,1986; Callon, Law, & Rip,1986).

Thus three word similarity matrixes were generated so social network analytical methods such as cohesion, centrality and community-detection (Blondel et.al., 2008) could be performed with a view to exploring the social and cognitive structure of Buddhist studies manifested in its published literature in respective languages. Specifically, the study seeks to answer the following interrelated research questions: firstly, are there recognizable branches or specialties in Buddhist studies? If so, what might these areas of research be? Cross-languages comparisons were also made to examine the similarities and differences of the intellectual structure in different language communities.

Results

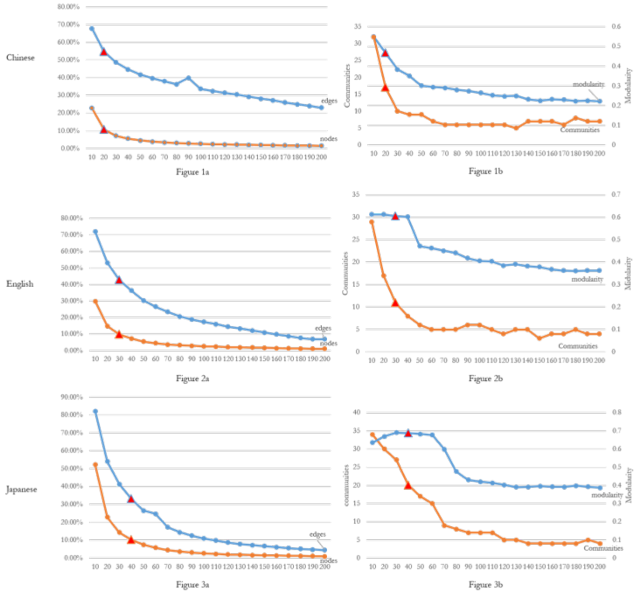
Descriptive data

Table 1 gives the numbers of items pertinent to various publication types analyzed in three languages in DLMB.

Language Publication Type	Chinese	English	Japanese
Journal Article	45,025	18,703	33,311
Book	12,644	13,604	8,869
Thesis and Dissertation	3,757	1,894	49
Research Paper	3,372	200	663
Proceeding Article	2,409	248	84
Journal Article; Book Review	158	2,201	377
Sound Recording	265	516	16
Serial	427	240	96
Reference Book	393	139	39
Audiovisual	45	304	16
Book Review	49	225	22
Internet Resource	59	207	0
Collected Papers	65	15	83
Others	15	33	0
E-Book	2	21	0
Book; Internet Resource	0	7	0
Book; Sound Recording	1	0	0
Internet Resource; Book Review	2	0	0
Internet Resource; Journal Article	0	2	0
Book Review; Internet Resource	0	1	0
Total	68,688	38,560	43,625

Table 1. Types of publications analyzed

Due to the enormous size of the networks, some sorts of filtering are required to make the groupings intelligible; node degrees was used as the filter as many of the little connected nodes tend to generate noises that impairs meaningful interpretation. To determine the proper threshold of node degree, one needs to consider three criteria: the quality of the clustering, the interpretability of the individual clusters, and the preservation of information. A high threshold would filter out large amount of nodes hence the greater loss of information and low modularity values. On the other hand, a low threshold would result in difficulty in interpreting individual clusters as they tend to lump together heterogeneous topics. Thus a trade-off needs to be made. We approached this matter by performing modularity analysis at different threshold levels so the values of their modularity, the resulting number of communities, as well as the size of the networks could be recorded. The following heuristics were used to select the proper thresholds: to preserve about 10 percent of the total nodes, to limit the number of communities from 10 to 20, and to preserve a high degree of modularity, which is commonly used as the indicator of clustering quality.



Figure(1-3). Node degree thresholds and resulting network attributes in three languages

Table 2 reports the thresholds and their corresponding attributes of the resulting networks.

	# of nodes	# of edge	Threshold	# of Communities	% of nodes	% of edges
Chinese	58,808	787,682	20	17	10.66	54.61
English	34,093	325,161	30	11	9.66	42.74
Japanese	75,728	859,469	40	20	10.09	33.04

Table 2. Descriptive statistics of the three networks

After filtering out lesser connected nodes, modularity maximizing community detection method was then performed to identify the subdomains in each language network. A two-stage approach was adopted here. As some of the communities resulting from the first-round of clustering can still be very broad and heterogeneous, a second modularity analysis was performed on these relative "super" clusters (i.e. clusters with more than 400 nodes), the joint results produce a two-levels hierarchical structure. Three experts in Buddhist studies were then interviewed to help us interpret the clusters (See Figure 4 and 5).

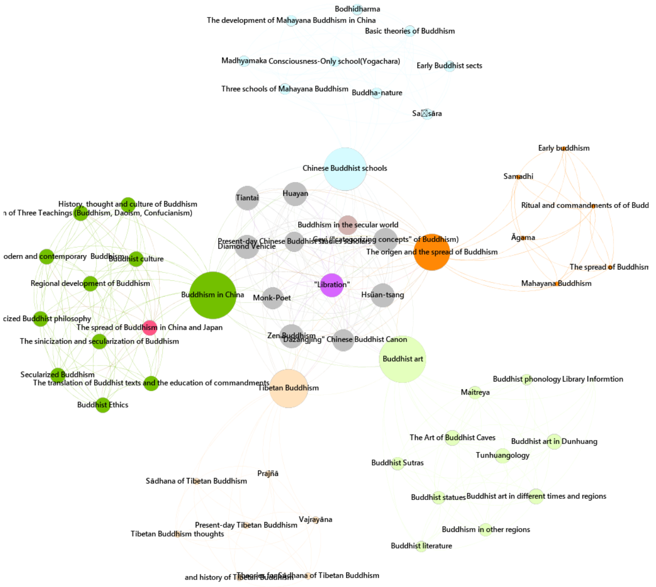

Bodhidharma
The development of Mahayana Buddhism in China
Basic theories of Buddhism
Madhyamaka Consciousness-Only school(Yogachara)
Early Buddhist sects
Three schools of Mahayana Buddhism
Buddha-nature
Early buddhism
Chinese Buddhist schools
Samadhi
Tiantai
Huayan
History, thought and culture of Buddhism
Ritual and commandments of of Budd
Buddhism in the secular world
Diamond Vehicle
Present-day Chinese Buddhist
Āgama
The origen and the spread of Buddhism
Buddhist culture
The spread of Buddhism
Regional development of Buddhism
"Libration"
Mahayana Buddhism
Buddhism in China
Monk-Poet
Hsüan-tsang
The spread of Buddhism in China and Japan
The sinicization and secularization of Buddhism
Zen Buddhism
Chinese Buddhist Canon
Buddhist art
Secularized Buddhism
The translation of Buddhist texts and the education of commandments
Tibetan Buddhism
Buddhist phonology Library Informtion
Buddhist Ethics
Maitreya
The Art of Buddhist Caves
Buddhist art in Dunhuang
Tunhuangology
Buddhist Sutras
Sādhana of Tibetan Buddhism
Prajñā
Buddhist statues
Buddhist art in different times and regions
Vajrayāna
Present-day Tibetan Buddhism
Tibetan Buddhism thoughts
Buddhism in other regions
Buddhist literature
Sādhana of Tibetan Buddhism

Figure 4. Visualization of intellectual structure in Chinse Buddhist studies.

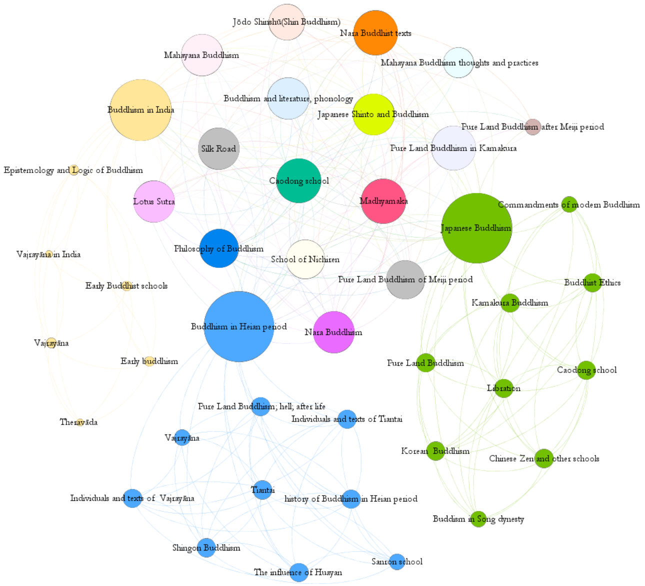

Jōdo Shinshū(Shin Buddhism)
Nara Buddhist texts
Mahayana Buddhism
Mahayana Buddhism thoughts and practices
Buddhism and literature, phonology
Buddhism in India
Japanese Shinto and Buddhism
Pure Land Buddhism after Meiji period
Silk Road
Pure Land Buddhism in Kamakura
Epistemology and Logic of Buddhism
Caodong school
Lotus Sutra
Madhyamaka
Commandments of modern Buddhism
Japanese Buddhism
Vajrayāna in India
Philosophy of Buddhism
School of Nichiren
Early Buddhist schools
Pure Land Buddhism of Meiji period
Buddhist Ethics
Kamakura Buddhism
Buddhism in Heian period
Nara Buddhism
Vajrayāna
Early buddhism
Pure Land Buddhism
Caodong school
Libration
Theravāda
Pure Land Buddhism; hell, after life
Individuals and texts of Tiantai
Vajrayāna
Korean Buddhism
Chinese Zen and other schools
Individuals and texts of Vajrayāna
Tiantai
history of Buddhism in Heian period
Buddhism in Song dynasty
Shingon Buddhism
Sanron school
The influence of Huayan

Figure 5. Visualization of intellectual structure in Japanese Buddhist studies.

In this study we utilized co-words networks to visually represent the domain of Buddhist studies. A heuristic was proposed to help select the proper threshold in order to filter less significant keywords. A two-stage clustering approach was adopted, which arguably provides a finer representation of the intellectual structure of the domain. Further analysis will be done, with the help of domain experts, to compare the differences in the intellectual structure reflects in three language communities.

Bibliography

Bhattacharya, S. and Basu, P. (1998), “Mapping a research area at the micro level using co-word analysis”, *Scientometrics,* 43(3), pp. 359-372.

Blondel, V., Guillaume, J., Lambiotte, R. and Lefebvre, E. (2008), “Fast unfolding of communities in large networks”, *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), pp. 10008.

Buitelaar, P., Bordea, G., & Coughlan, B. (2014), “Hot Topics and Schisms in NLP: Community and Trend Analysis with Saffron on ACL and LREC Proceedings”. In *9th Edition of Language Resources and Evaluation Conference (LREC2014)*.

Cabezón, J. I. (1995), “Buddhist Studies as a Discipline and the Role of Theory”, *Journal of the International Association of Buddhist Studies*, 18(2), pp. 231-268.

Callon, M., Courtial, J. and Laville, F. (1991), “Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry”, *Scientometrics*, 22(1), pp. 155-205.

Callon, M., Law, J., & Rip, A. (1986), “Qualitative scientometrics”, in *Mapping the dynamics of science and technology*, Palgrave Macmillan UK, pp. 103-123.

Courtial, J. (1994), “A coword analysis of scientometrics”. *Scientometrics*, 31(3), pp. 251-260.

Courtial, J. P. (1986), “Technical issues and developments in methodology”, in *Mapping the Dynamics of Science and Technology*, Palgrave Macmillan UK, pp. 189-210.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001), “Bibliometric cartography of information retrieval research by using co-word analysis”. *Information processing & management*, 37(6), pp. 817-842.

He, Q. (1999), “Knowledge discovery through co-word analysis”, *Library trends*, 48(1), pp. 133-133.

Leydesdroff, L. (1989), “Words and co-words as indicators of intellectual organization”, *Research policy*, *18*(4), pp. 209-223.

Peters, H. P. F., & van Raan, A. F. (1993), “Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling”, *Research Policy*, 22(1), pp. 23-45.

Tijssen, R. J. (1992), “A quantitative assessment of interdisciplinary structures in science and technology: co-classification analysis of energy research”, *Research policy*, 21(1), pp. 27-44.

Reading Ancient Scripts: Investigating the Human Visual System for Artificial Intelligence in Palaeography

Ségolène Tarte
segolene.tarte@oerc.ox.ac.uk
University of Oxford, United Kingdom

Rachel Mairs
r.mairs@reading.ac.uk
University of Reading, United Kingdom

Mihaela Duta
mihaela.duta@psy.ox.ac.uk
University of Oxford, United Kingdom

Chrystalina Antoniades
chrystalina.antoniades@ndcn.ox.ac.uk
University of Oxford, United Kingdom

With the renewed vitality of research in Artificial Intelligence, thanks in particular to the continued development of neural network techniques for machine learning, computer vision technologies developed for handwriting recognition offer innovative ways of conducting research in palaeography (Brusuelas, 2016; Hassner et al 2014; Muzurelle, 2011; Stutzmann, 2015)

In this context where artificial intelligence often endeavours to replace human intelligence, or at least to emulate it, we are undertaking to understand better what it is that human intelligence does when reading ancient handwritten scripts. Ultimately, our ambition is to nudge artificial intelligence for palaeography to be intelligent enough to identify where human intelligence is still superior to machine intelligence (and therefore better left the upper hand) and where researchers can benefit from algorithmic support.

Handwriting recognition is a challenging task – both for humans and machines – because handwritten scripts inherently straddle two interlinked spheres of intelligence: that of visual shapes, and that of semantics.

This work builds on previous research (Terras, 2006; Youtie, 1963) that has identified six strands of human strategies in palaeography through ethnographic analysis, the results of which were cross-referenced with the cognitive sciences literature (Tarte, 2014). These strands were: visual perception, aural feedback, motor feedback, semantic memory, structural knowledge acquisition, and creativity; all continuously interacting with and feeding back into each other to some degree. In this project, we concentrate on the task of reading ancient handwriting – one of the many aspects of palaeographical research, whether it is concerned with mediaeval scripts or with more ancient scripts.

In this paper, we present some findings and observations made in the process of designing experiments to investigate some of the mechanisms underlying handwriting recognition in a palaeographical context; preliminary results from the experiments themselves are forthcoming.

To explore in depth how humans handle the variability of the shapes of signs in a given script, our experiments aim to bridge between traditional ethnographic methodologies, geared towards the gathering of qualitative data, and cognitive sciences methodologies, geared towards the gathering of quantitative data. The script of choice was Demotic, a script of the Late Egyptian language and whose use spanned ten centuries, therefore displaying a large variability in shapes. The team of scholars involved in designing and conducting our experiments was made of: an Egyptologist/Classicist, an Ethnographer, a Neuroscientist, and a Computer Scientist. Many of the observations reported here stem from the epistemological encounters of very different traditions of research; they emerged through the interdisciplinary conversations that took place in the process of designing the experiments.

The outcome of these conversations was the following experimental setup, building on the principles of the protocols outlined by Althaus and Plunkett (2015) and Longcamp et al (2008).

Experiment

Volunteers are invited to two experimental sessions that take place in a library setting, where they interact with a tablet computer using a stylus. The first session is a learning session and the second is a delayed recognition session. The sessions take place at least one week apart.

During the learning session volunteers learn to recognise 5 Demotic signs (target signs). This session comprises of a familiarisation phase followed by an immediate recognition phase. The familiarisation phase can comprise of one of the following three familiarisation conditions:

- static passive familiarisation – 3-second repeated presentation of each sign
- static active familiarisation – repeated presentation of each sign with time for the volunteer to draw the sign on the tablet using the stylus
- dynamic familiarisation – 3-second repeated presentation of movies depicting the drawing of each sign

During the familiarisation phase each sign is presented 8 times, twice in each of 4 distinct hands. The presentation order is pseudo-randomize to ensure that signs don't appear twice in a row. Each volunteer is assigned randomly to one of the three familiarisation conditions. The familiarisation phase is followed by a 2-step immediate recognition phase. In the first recognition phase, pairs of Demotic signs are presented; each pair is made of one target sign and one distractor sign (the distractors are also Demotic signs), and in the second recognition phase words containing the target sign are presented.

The delayed recognition session comprises of three phases: a delayed recognition, a repeated familiarisation and a re-enforced delayed recognition. The two delayed recognition phases are similar to the immediate recognition phase, while the repeated familiarisation is similar to the familiarisation of the learning session.

At the first session, all volunteers are given a short video introduction that aims to prime them towards scripts and writing before starting the experiment; the second session ends with a debrief where volunteers are given the freedom to ask questions about the tasks or scripts to the experimenter.

Negotiations

In the process of designing this experimental setup, a number of fascinating questions emerged due to the intrinsic multidisciplinary nature of the project. The main interdisciplinary negotiations that took place can be summarised thematically as revolving around: the definition of an alphabet; the nature of script and materiality; and biases and mixed methodologies.

Alphabet

As the researchers and the volunteers evolve in an environment where the default script is alphabetic, we decided to choose an alphabetic script. The choice of Demotic can therefore be questioned: Demotic is not an alphabetic script, even if most of its signs have a phonetic value. We wanted however to use real data – as opposed to an invented alphabet – so was there any context in which Demotic might have been used as an alphabet? From the Ptolemaic period onwards, the frequency of Greek names in Egypt is higher, and therefore documents bearing Greek names resort to transliterating them into Demotic. Although there was no received orthography for those transliterations, it becomes more acceptable, in this specific context, to consider Demotic signs as alphabetic signs (for the purpose of this experiment, determinatives, when present, were regarded as not part of the word).

In turn, this deliberate choice also made a search for images of signs written in different hands easier. By querying http://www.trismegistos.org/ (Brouz and Depauw, 2015) for a list of Greek personal names in Demotic documents, and cross-referencing the results of this search with those from a query on papyri.info for all papyri in Demotic that have accompanying digital images, it was possible to identify and isolate images of Greek personal names written in Demotic on papyri.

Nature of script and materiality

The question of the homogeneity of the signals presented to the volunteers is important in the cognitive sciences. However, isolating the signs from their support means possibly removing information (e.g. faint ink on more salient fibres of a papyrus, degraded or stained papyri). So in an effort to present realistic data, we have decided to use greyscale images, to crop the images of words/signs and to simply uniformize the overall look of the images of words by aligning their histograms; we have also endeavoured to present all signs in such a way that they have a similar size (so some scaling was performed).

A further question was that of the phonetic dimension of the script. From a cognitive sciences perspective, as the focus is on the visual, it didn't make sense to add a phonetic element to the script at this stage. Only when it is better understood how the visual processes handle variability in shape, can it be envisaged to add a multisensory layer of complexity.

This raises some intriguing questions with regard to the nature of a script. In particular, isn't the essence of an alphabetic script to be a notation system for phonetic word utterances? What does the removal of its phonetic dimension entail for the script (regardless of whether the existing phonetic dimension is a modern convention or an actuality)? Are we denaturing the script by presenting its signs stripped of their phonetic values?

Biases and mixed methods

Negotiating between the highly-controlled design of experiments in the cognitive sciences and the naturalistic settings and exchanges of ethnography proved complex. The main concern in such endeavours is to not compromise the validity of the gathered data with respect to the frameworks of the traditional disciplines. Free exchanges between volunteer and experimenter have the potential to bias the overall results for the cognitive sciences, as the bias is uncontrolled; it was therefore important for any such exchanges to happen only after all data of the controlled variables are collected.

Even then, leaving space for free exchange within the confines of the experiment worried the cognitive scientists who feared to open the door to highly irrelevant questions or even somewhat leftfield questions (e.g. "Can you see my gift of clairvoyance?") It appeared that such questions might be favoured by the running of such experiments in medicalised environments (hospital, psychology department), so we ran our experiments at the library instead, thereby establishing an environment resonant with our overarching theme of reading.

The process of designing these experiments as well as the forthcoming results have some bearing on the understanding of expertise of course, but also on how one might proceed when faced with a large corpus of unedited textual artefacts: Can crowdsourcing approaches be specifically geared towards the strength of the human visual system? Can algorithmic approaches palliate the weaknesses of the human visual system?

Acknowledgments

This project is funded by the John Fell Oxford University Press (OUP) Research Fund.

Bibliography

Althaus, N., and Plunkett, K. (2015). "Timing matters: The impact of label synchrony on infant categorization." *Cognition*, 139:1- 9.

Broux, Y., and Depauw, M. (2015). "Developing onomastic gazetteers and prosopographies for the ancient world through named entity recognition and graph visualization: Some examples from trismegistos people." In L. Aiello and D. McFarland (eds), *Social Informatics. SocInfo 2014 International Workshops, GMC and Histinformatics*, Lecture Notes in Computer Science 8852, Springer, pp 304-13.

Brusuelas, J. (2016). "Engaging Greek: Ancient lives". In G. Bodard and M. Romanello, (eds), *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement*, Ubiquity Press, London, pp 187-204.

Hassner, T., Sablatnig, R., Stutzmann, D., and Tarte, S. (2014). *Digital Palaeography: New Machines and Old Texts* (Dagstuhl Seminar 14302). Dagstuhl Reports, 4(7):112-34.

Longcamp, M., Boucard, C., Gilhodes, J.-C., Anton, J.-L., Roth, M., Nazarian, B., and Velay, J.-L. (2008). "Learning through hand- or typewriting influences visual recognition of new graphic shapes: Behavioral and functional imaging evidence." *Journal of Cognitive Neuroscience*, 20(5):802–815.

Muzerelle, D. (2011). "À la recherche d'algorithmes experts en écritures médiévales." *Gazette du livre médiéval*, 56–57:5–20.

Stutzmann, D. (2015). "Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol." *Digital Medievalist*, 10.

Terras, M. (2006) *Image to Interpretation. An Intelligent System to Aid Historians in Reading the Vindolanda Texts*. Oxford University Press, Oxford.

Tarte, S. (2014). "Interpreting textual artefacts: Cognitive insights into expert practices." In C. Mills, M. Pidd, and E. Ward, (eds), *Proceedings of the Digital Humanities Congress 2012, Studies in the Digital Humanities*. Sheffield: HRI Online Publications.

Youtie, H. C. (1963). "The papyrologist: artificer of fact." *Greek, Roman and Byzantine Studies*, 4(1):19-33.

Generous and Generative Communities for the Digital Humanities with the Digital Library of the Caribbean and Caribbean Studies

Laurie N. Taylor
laurien@ufl.edu
University of Florida, United States of America

Introduction

The digital humanities offers the promise and potential for global, multilingual, and multicultural collaborations as enabled by digital technologies that deliver public-facing scholarship that enriches and expands research communities. Realizing that potential requires approaches that embrace complexity, spanning the technological, social, procedural, and community concerns. In 2016-2017, the Digital Library of the Caribbean (dLOC) is undertaking a research project to study and advance support for the Digital Humanities specifically with Caribbean Studies, focusing on leveraging technologies to support public-facing scholarship, open access to research and results, enabling digital humanities research publications with technologies that meet access needs, and collaboration among scholars and communities. This presentation provides background on dLOC as a digital library, dLOC's evolution into a community connecting place and platform for digital humanities and Caribbean Studies, and the opportunities and needs for the digital humanities for enlarging scholarly and worldwide community access to Caribbean Studies.

About dLOC

The Digital Library of the Caribbean (dLOC) is a unique, open access, collaborative, international, multi-lingual digital library for resources from and about the Caribbean and circum-Caribbean, providing access and ensuring preservation for Caribbean materials (digitized and born-digital) and public-facing scholarship.

dLOC focuses on how a community of practice can best create a digital library in terms of contents, functionality, and robust governance for inclusivity and diversity. In dLOC's governance and operational model, partner institutions agree to shared goals and processes following a governance structure for: inclusive and distributed collection development where partners select materials, permissions-based infrastructure (partners retain all rights to materials), functional hubs, decentralized/local digitalization and digital curation, collaborative activities to develop the community of practice and increase capacity through collaboration. For all of this work—spanning tools for digitalization, online digital library functionality for patrons and users, tools for advanced querying and mining, and tools for automating reporting and user management which are necessary for operations at scale and reports to different governmental and academic entities—dLOC's technologies have been defined by and created in collaboration with partner institutions and scholars. Thus, over the past 12 years, dLOC has developed through collaboration into as a socio-technical (people, policies, communities, technologies) platform supporting collaboration among partner institutions, developing and enhancing communities of practice, and building intellectual infrastructure.

dLOC, Caribbean Studies, and Digital Humanities

dLOC originally focused on building content, and has grown into one of the largest open access collections of Caribbean materials with over 2.5 million pages of content, over 40 partner institutions (universities, colleges, libraries, archives, museums, government agencies, NGOs, publishers, and scholarly societies, as well as many contributing scholars and private collectors), and over 3 million views each month. With this growth, the dLOC Executive Committee, Scholarly Advisory Board, and full community recognized the need to emphasize new stages of development, specifically focused on further engaging scholars in digital humanities practices that build upon dLOC's commitment to access, preservation, the production of public-facing scholarship, and engaging across institutions and communities to further the community of practice as a constellation of communities of practice where all involved are leveraging the affordances of technology to further public humanities and interdisciplinary aspects of modern scholarship.

In 2016, the dLOC Executive Committee charged the dLOC Digital Scholarship Director with undertaking research into next steps for dLOC's socio-technical development (people, policies, technologies, communities) in relation to new opportunities with the digital humanities. dLOC currently supports the digital humanities in many forms, including curated materials and collections, digital humanities exhibits, pedagogical resources, teaching guides, and supporting faculty in developing online research and teaching materials.

dLOC's work in the digital humanities grapples directly with questions of access in the digital age. For example, one of dLOC's digital humanities projects is *Haiti: An Island Luminous*. *Haiti: An Island Luminous* began when then-PhD student Adam Silvia recognized the importance of materials about and from Haiti in dLOC, both for the significance of each item and for the sheer scale of materials available. Many of the materials in dLOC were not known to exist in the world, before being located by partner institutions and then digitized. Once the materials became available, the necessary scholarly and information ecosystems to link and cite the items was not in place. Indeed, scholars would be unlikely to search for items that were believed to have been lost to history. To create the necessary community connections, Silvia contacted members of the dLOC team and began collaborating to develop what was planned as an online exhibit to showcase materials. As the project began, the scope continued to grow. In its final release, *Haiti: An Island Luminous* is a curated edited online collection with contributions from over 100 of the top Haitian Studies scholars. *Haiti: An Island Luminous* is in English, French, and Kreyòl, and has been taught in schools in Florida and Haiti. Developed first online, the site is being developed for offline use on standalone kiosk/tablet installations, to meet access needs despite limited connectivity. Discussions are underway on the possibility of a print version to meet needs for limited online and electrical access, with the print version to include all pages from the curated edited collection online, samples of items, and with an accompanying USB drive.

Another example of dLOC's digital humanities work is the course "Panama Silver, Asian Gold: Migration, Money, and the Making of the Modern Caribbean." The course focused on the lesser studied Asian and Indian indenture in the Caribbean and was a Distributed Online Collaborative Course (DOCC), taught in multiple semesters and multiple iterations, most recently with the teaching team comprised of faculty, librarians, and archivists at the University of the West Indies, Cave Hill Barbados, University of Miami, University of Florida, and Amherst College. This collaboratively developed and taught course engaged the teaching team and students across all of the campuses in the use of materials in dLOC, other digital resources, and not-yet-digital resources held at each of the institutions to create new works of digital scholarship with students from different campuses working collaboratively. The course is underway for revision to focus on migration and mobility, to further expand the potential collaborators and fields engaged in this interdisciplinary collaborative course.

Both the DOCC series and *Haiti: An Island Luminous* present real examples of imagined possibilities for expanding Caribbean Studies in the digital age for creating public-facing scholarship, enhancing access to scholarly works, opportunities from the digital humanities in pedagogy and academic curricula, and the potentials for access as enabled by collaboration among scholars and communities.

Expanding Access through dLOC for Caribbean Studies and Digital Humanities

Caribbean Studies is an interdisciplinary field with broad connections across languages and cultures. dLOC takes its definition of the Caribbean from the Association of Caribbean University, Research, and Institutional Libraries, which defines the Caribbean as “the area of the Caribbean archipelago, the mainland countries including the Guianas, and the states of the United States which border on the Caribbean Sea or Gulf of Mexico.” Even this is immediately expanded with diasporic connections. dLOC’s success as a digital library is made possible by the recognition that the Caribbean exceeds the boundaries for any specific geographic area, language, or field of study, and by the generous framing for Caribbean Studies with the respect and support for interdisciplinary and diasporic connections.

dLOC was founded to meet the needs for preservation and access as a first step in supporting expanding the field of Caribbean Studies. As both the series of DOCCs and *Haiti: An Island Luminous* demonstrate, simply creating online access to materials is only part of the equation. As Adam J. Banks explains in *Race, Rhetoric, and Technology: Searching for Higher Ground*, access includes material access, which is met with materials being online, at least for those with online access and the functional access to be able to find and use materials. The digital humanities offers ways of conceiving and creating the means for experiential access, where materials are relevant for people’s lives, and transformative access, where there is “a genuine inclusion in technologies and the networks of power that help determine what they become, but never merely for the sake of inclusion” (45).

This presentation will review methodologies and results from the 2016-2017 research on the digital humanities and Caribbean Studies. The research includes site visits to the US Virgin Islands, Jamaica, Barbados, Curaçao, Trinidad, and Leiden (for the Dutch Caribbean Archives) to discuss current activities, projects, terminology, and framing to enable connections across communities. The site visits include engagement with scholars from multiple fields, museum professionals, librarians, archivists, governmental representatives, educators, and others. All materials from the research visits are shared with communities engaged and connected with Caribbean Studies through conferences and various means. This grounded approach allows for identification, recognition, and connection of digital humanities work being done, even when not previously labeled as such. The presentation will include findings on how dLOC is enabling responses and activities to meet the needs for material through transformational access through technologies, procedures, communities, and technologies.

Accessing Russian Culture Online: The scope of digitisation in museums across Russia

Melissa Terras
author.email@domain.com
University College London, United Kingdom

Inna Kizhner
inna.kizhner@gmail.com
Siberian Federal University, Russia

Maxim Rumyantsev
m-rumyantsev@yandex.ru
Siberian Federal University, Russia

Kristina Sycheva
kristina.sycheva.2012@mail.ru
Siberian Federal University, Russia

Although the rate and coverage of digitization throughout Europe is monitored and understood (Europeana, 2016; Minerva EC, 2016, Navarette 2015) there has been little work done on understanding the reach of digitization across Russia. In this paper, we build on previous work (Kizhner, Terras, Rumyantsev, 2016) by using Russian Ministry of Culture statistics to calculate the percentage of museum collections that have been digitized across Russia. We show regional variations and demonstrate that although many Russian museums have digitisation programs, this is not carried out to the same extent as across Europe. We suggest that studying non-European digitization practices can lead to further understanding of the digital canon upon which analysis of culture is based (Limb, 2007; Warwick et al, 2008; Price, 2009; Earhart, 2012).

Digital visual collections from national and regional museums can be a rich source of data for digital humanities but the first step in these studies might be to discover the existence of digital images on national levels. Do non-European digital collections exist? Do non-European museums provide the same amount of data as European museums? Is this amount equal for the centre of the country and for provinces? How many images are posted online?

This paper aims to find out the current scale of digitization in Russian museums. We discuss Russian digitization as an example of real life messy collection of data where the situation is hardly related to «post-modernist utopia» (Sartori 2016) or the wishful thinking of moving images across interfaces (Robinson, 2013; Kizhner, Stankevich, Terras, Rumyantsev 2016), all of them quite distant perspectives.

Starting from the 1970s, the rationale for museum digitisation practices in Russia was quite similar to that in many other countries. It was informed by a need for information and collection management so that museum objects would not be lost and were properly conserved (Navarette, 2014; Sher, 1983; Williams, 1987; Chenhall and Vance, 1987). Russian government policy related to that need from 2008 onwards was aimed at building The National Catalogue of Museum Objects posted online (Ministry of Culture of the Russian Federation, 2016) and currently including images for 1,2 million museum objects, 1,5% of total number of Russian museum objects (slightly over 80 million), and 2% of the collection of unique objects (about 60 million).

The National Catalogue is an initial access point in finding out the scale of museum digitization in various parts of the country including its remote regions. Our previous paper (Kizhner, Terras, Rumyantsev, 2016) demonstrated preliminary results of a survey estimating the percentage of digital images for Russian museum collections. The study also included web site exploration results on the percentage of museum collections posted online. However, it only covered 1,2% museums in the country for the percentage of digitized images and 6% for the images posted online and its results gave initial estimates.

The present paper studies the percentage of digital images through the statistical reports submitted to the Ministry of Culture from 2,367 museums in 2015. The annual statistical reports are mandatory for all museums reporting to local municipalities, regional administrations and the RF Ministry of Culture, in fact for all non-private and non-corporate museums. We provide the average results for the country and and the average results for its 8 major geographical regions.

Methodology

The data of the RF museums' statistical reports for 2015 were received from the RF Ministry of Culture in summer 2016. Museums return these mandatory reports in January and provide statistical data for the preceding year. The data were related to 2,635 museums from every region of the Russian Federation.

The data were received as an Excel spreadsheet from which the parts were removed that were not related to the digitization of museum objects and the data on galleries that were for temporary display and did not store any objects. This left us with 2,367 museums. The data in the spreadsheet were sorted on the total number of objects for every museum, the number of unique objects, the number of database records with digital images, and the number of images posted online.

Russian museum collections tend to consist of two parts: the main collection of unique objects and a smaller 'research collection' including duplicates and supporting documentation. While the total number of objects in Russian museum collections slightly exceed 80 million objects, the number of unique objects is 20 million fewer and equals 60 million objects. The results of statistical surveys obtained for the study reported the number of digitized objects as related to the total number of objects in a museum including their 'research collections'. This did not create a methodological problem when comparing the results with those from the Enumerate project which is a study of the outage of digitisation across Europe, funded by the European Union which happened between 2011 and 2015 (Europeana, 2016) where the survey asked to provide the percentage of digital images for museums' analogue collections.

The percentage of digitized objects and the percentage of objects with images posted online was calculated for each geographical region of the Russian Federation and mapped to show the differences. The total percentage for the country was also assessed.

Results

The percentage of digital images as related to the total number of museum objects across Russia was 13,5%. The percentage related to the number of objects in the main collection (roughly corresponding to the number of unique objects) was 18%.

The percentage of images posted online as related to the total number of objects was 1,5%, this figure was somewhat larger if compared with the number of objects in the main collection (2%).

An interesting and unexpected result was the difference between the scale of digitization in two major cities, Moscow and Saint Petersburg (Table 1). The percentage of objects with digital images was much higher than the average across Russia in Saint Petersburg and somewhat lower than the average in Moscow. The scale of digitization across major geographical regions varied between the minimum of 6% in the Far East and the maximum of 25% in the regions adjacent to Saint Petersburg (Figure 1, Table 1).

Interestingly, the percentage of images posted online was slightly lower than the average across Russia for museums in Moscow and twice lower than the average across Russia in Saint Petersburg (Table 2).

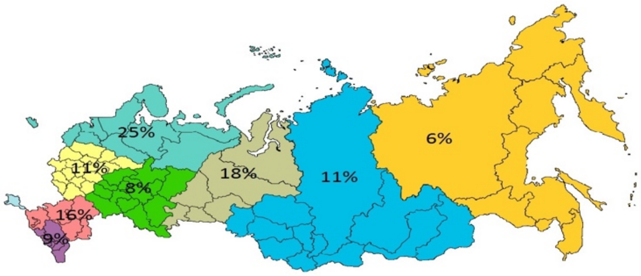


Figure 1. The percentage of the analogue collections digitally reproduced across geographical regions

Places	% as related to the total number of objects	% for the main collection (unique objects)
Saint Petersburg	36	44
Moscow	10	12
The average across Russia	13,5	18
Centre (Central Federal District)	11	14
North-West (North-Western Federal District)	25	33
Southern Federal District	16	23
Caucasus (North-Caucasian Federal District)	9	12
Volga Federal District	8	11
Ural Federal District	18	25
Siberian Federal District	11	16
Far Eastern Federal District	6	8

Table 1. The percentage of the analogue collections digitally reproduced in the museums of Saint Petersburg, Moscow, and across Russia

Places	% as related to the total number of objects	% for the main collection (unique objects)
Saint Petersburg	0,9	1,1
Moscow	1,2	1,5
The average across Russia	1,5	2

Table 2. The percentage of digital images posted online

Discussion

Our findings demonstrate that digital collections in Russian museums do exist across the country but we cannot say that their online display is representative enough to cover the culture, considering the variety in geography and ethnography.

We can roughly confirm our previous results on the percentage of digitized images (Kizhner, Terras, Rumyantsev, 2016) to be around 18% as our present data show the level of digitization to be in the range of 13,5 - 18%. However, our previous results might have a sampling bias as the museums answering the questions of the survey could be interested in digitization per se and work towards obtaining more financial and administrative support to keep it going. This might result in a higher percentage. Even so, the results of the present study can only demonstrate a range of digitization scale as the percentage of images related to the total number of objects may include a significant number of duplicates and supporting materials (e.g. library books).

This is not an obstacle to comparing our data with those from the Enumerate project 'which aimed to survey the extent of digitization across Europe' (Europeana, 2016) where the survey questions were about the percentage of the analogue collection digitally reproduced, but given the range of 13,5 - 18% we can say that the results for 2015 are much lower than the results of the Enumerate project for 2015 when the percentage of digitized collections in European museums was 31%. However, the results for Saint Petersburg collections are higher than the European average

(Table 1).

The percentage of images available online across Russia as related to the analogue collection is 1,5 - 2% which is lower than the percentage reported by the Enumerate project (24% of digital collections and 7,5% of European analogue collections). However, the Enumerate results included digital collections and digitally born objects available online, which complicates the comparison (Europeana, 2016).

Recent criticism of digitization without proper contribution to building knowledge in the humanities (Hitchcock, 2013, Gregory et al., 2016) requires developing these studies further towards exploring how Russian museum web sites arrange images for searching and browsing and developing projects discussing the issues of open access and digital canon.

This paper opens up the space for studying Russian digital collections on a national scale. It also reports on the results of looking at the scale of digitization for major geographical regions within Russia, and it will discuss the results of calculating simple correlations of digitization percentage with population density, the level of education, funding, and the number of museum goers in the regions. By doing so we can challenge the concept of the digital canon, and ask difficult questions regarding which types of culture are being digitized and made available worldwide.

Bibliography

Chenhall, R., and Vance, D., (2010) 'The World of (Almost) Unique Objects'. In *Museums in a Digital Age,* ed. Ross Parry, London, New York: Routledge, pp. 39-48.

Earhart, A., (2012). Can Information Be Unfettered? Race and the New Digital Humanities Canon. Debates in the Digital Humanities, pp.309-318.

Europeana (2016), Enumerate Observatory. http://pro.europeana.eu/enumerate/

Gregory, I., Atkinson, P., Hardie, A., Joulain-Jay, A., Kershaw, D., Porter, C., Rayson, P., Rupp, CJ. (2016). From Digital Resources to Historical Scholarship with the British Library 19th Century Newspaper Collection, *Journal of Siberian Federal University. Humanities and Social Sciences*, 9 (4), pp. 994-1006.

Hitchcock T. (2013) "Confronting the digital or how academic history writing lost the plot" *Cultural and Social History*. 10, pp. 9-23.

Kizhner, I., Stankevich, J., Terras, M., Rumyantsev, M., (2016). Licensing Images from Russian Museums for an Academic Project within Russian Legislation". *Special Interest Group Au-*

dioVisual Material in Digital Humanities Workshop , Digital Humanities 2016, Krakow. https://avindhsig.wordpress.com/workshop-2016-krakow/accepted-abstracts/

Kizhner, I., Terras, M., Rumyantsev, M. (2016). Museum Digitization Practices Across Russia: Survey and Web Site Exploration Results. In *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, pp. 600-602.

Limb, P., (2007). The politics of digital" reform and revolution": towards mainstreaming and African control of African digitisation. Innovation, 2007 (34).

Minerva EC, (2016). http://www.minervaeurope.org/home.htm

Ministry of Culture of the Russian Federation (2016), *The National Catalogue of the RF Museum Collections*, http://goskatalog.ru/portal/#/

Navarette, T., (2014). *A History of Digitization: Dutch Museums.* Ph.D. diss., University of Amsterdam. http://catalogus.boekman.nl/pub/P14-0752.pdf

Navarette, T., (2015). 'Benefits of Collaborative Digitization Projects in Europe', Les Cahiers Numerique, 2015/1, (Vol. 11). http://www.cairn.info/revue-les-cahiers-du-numerique-2015-1-page-41.htm

Price, K.M., (2009). Digital Scholarship, Economics, and the American Literary Canon. Literature Compass, 6(2), pp.274-290.

Sartori, A. (2015). "Towards an Intellectual History of Digitisation: Myths,Dystopias, and Discursive Shifts in Museum Computing". *Digital Scholarship in the Humanities*, March 2015.

Sher, Y., (1983). 'Preface from the Editor of the Russian Edition'. In Chenhall, R. *Museum Cataloging in the Computing Age,* eds. Yury Aseev and Yakov Sher, Moscow: Mir, pp. 7-17. In Russian.

Warwick, C., Terras, M., Huntington, P., and Pappa, N. (2008). 'If you build it will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data." Literary and Linguistic Computing 23, no. 1 (2008): 85-102.

Williams, D. (2010). 'A Brief History of Museum Computerization'. In *Museums in a Digital Age,* ed. Ross Parry, London, New York: Routledge, pp. 15-22

Personæ: A Character-Visualisation Tool for Dramatic Texts

Justin Tonra
justin.tonra@nuigalway.ie
National University of Ireland Galway, Ireland

David Kelly
author.email@domain.com
National University of Ireland Galway, Ireland

Lindsay Reid
lindsay.reid@nuigalway.ie
National University of Ireland Galway, Ireland

Introduction

This paper explores the development of the *Personæ* tool (code available on Github), an interactive resource for exploring patterns of speeches by and mentions of characters in dramatic texts. Initially developed to examine works by Shakespeare, the tool has broad application to dramatic texts.

Visualising the frequency, extent, and position of dialogue relating to a particular character presents users with a simple and immediate measure of that character's prominence within the play. The *Personæ* tool enables users to select and visualise individual characters' involvement, producing a novel means of exploring large-scale structural, narrative, or character-focused patterns within the text.

The tool is intended to facilitate character-based analysis and reveal structural patterns at the scale of the play. The tool was conceived with exploratory potential in mind, and is designed to allow users to customise the visualisation according to their particular interests or to follow a more speculative and disinterested reading of the play's character-based features.

This deliberate aim emerged from the heuristic development process described below, and a desire to produce an extensible exploratory tool for dramatic texts. From an initial focus on using digital tools to visualise the tangling and disentangling of character names and identities in *The Comedy of Errors*, our interest broadened into exploring the potential for using character data to visualise larger structural and narrative patterns.

We were also motivated by the use of network analysis and visualisation for scholarship on Shakespearean and other literary texts, including work by Yose et al (2016), Grandjean (2015), Moretti (2011), and Stiller, et al. (2003). These analyses are similarly character-based and have yielded many interesting insights. But in the reduction of the textual data to nodes and edges (characters and their interactions), network analysis has obscured the temporal. The work of Xanthos, et al. (2016) maintains this temporal dimension, while exploring the dynamics of the character networks as they evolve. In contrast, by visualising the characters at the level of the play as a whole, we aim to preserve characters' locations within the space of the text, thereby enabling analysis of the dramatic time and structural duration of the play.

Tool Development

Tool development took part in two phases. First, the data was extracted and transformed into a suitable format. The user interface was then designed using an iterative process that enabled the exploration of various approaches to data presentation and interaction.

Data Preparation

The tool uses data contained in XML files provided in the New Variorum Shakespeare editions of *The Comedy of Errors* and *The Winter's Tale*.

Data was extracted using a custom-developed Python script which iterates through each play's XML file extracting character and name data, along with line number, scene and act identifiers. The data as output as JSON, which reduces the complexity of using it with the JavaScript-based user interface.

User Interface Design

The tool's web-based user interface (Figure 1) was developed using the open source Javascript library, D3 (Bostock et al, 2011). *Personæ* developed from a fixed and static visualisation of *The Comedy of Errors* to a more interactive and exploratory tool. In the heuristic spirit of the tool itself, we describe here its various iterations, the stages of its development, and the motivations for various changes to its design and functionality throughout the process.

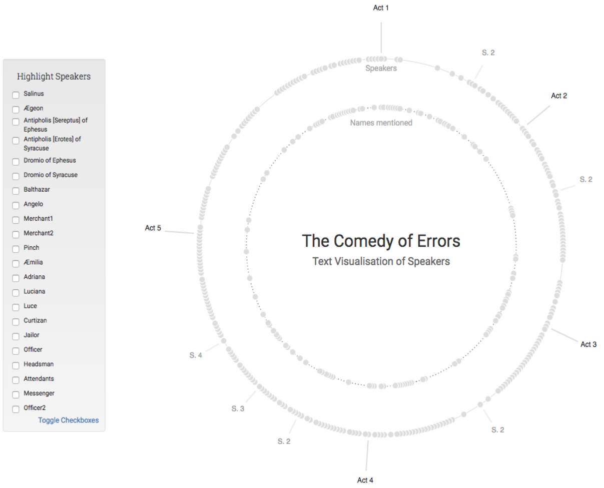


Figure 1: *Personæ* Character Visualisation Interface - From the First to the Second Iteration

Personæ's focus on character and temporal visualisation is present in the first iteration of the tool (Figure 2). Speeches and mentions are plotted along a timeline, with a tabular view switching between the five acts of the play. All speeches and mentions are colour-coded, resulting in some interesting patterns and densities at certain parts of the text, but lacking the facility for isolating chosen characters. In addition, the tabular view of the five acts lacked the desired holistic view of the entire play.

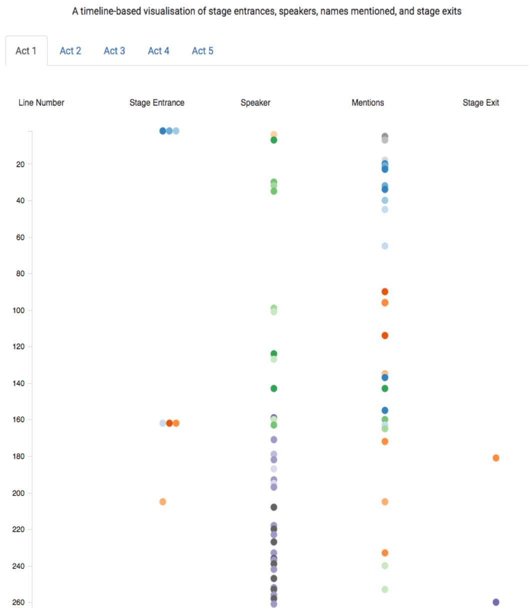


Figure 2: First iteration design

Expanding the Second Iteration

The second iteration of the tool adopted the circular layout of the tool to plot character involvement across the entire play, as shown in Figure 3. At this point, the tool was still static, and its focus on the two pairs of twin characters in *The Comedy of Errors* represented a desire to deploy visualisation for a particular exploratory purpose. The play operates on the basis of identity and confusion, as Antipholus and Dromio of Syracuse are mistaken for their Ephesian counterparts, and vice versa. Our aim was to plot the speeches of these four characters to see if the visualisation revealed any insights into how the identity question was introduced and managed at a structural level.

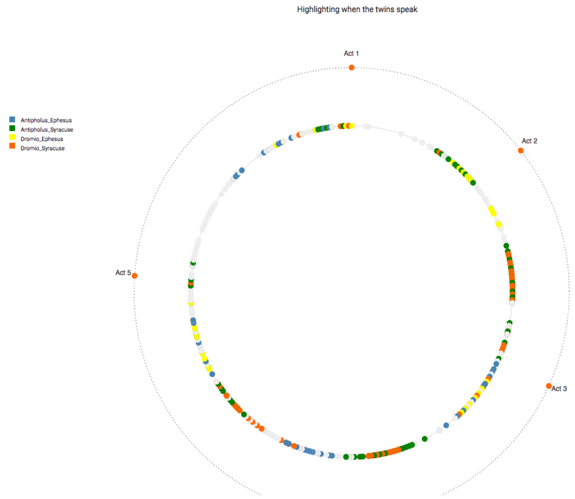


Figure 3: Second iteration design

Indeed, the visualisation presents several clustered scenes of engagement between the pairs of twins through which various errors and misunderstandings are played out. The tell-tale single appearance of Dromio of Syracuse's orange marker in Act 1, Scene 2 precisely represents the beginnings of the error and confusion: “What now? How chance thou art returned so soon?”

An additional avenue of exploration in the second iteration of the tool was the geographical mapping of locations mentioned in the play. *The Comedy of Errors* is known for including the only mention of America in Shakespeare’s plays, among several other placenames in its text. In some respects, this visualisation gives a false impression of *The Comedy of Errors* as a worldly play. While eighteen locations are mentioned in the text, several of these are ironically located by Dromio of Syracuse on Nell the kitchen-maid’s body, because “she is sphericall, like a globe: I could find out / Countries in her” (Act 3, Scene 2).

The final interface

As useful as this view of the play proved, we felt at this point that a more dynamic and interactive interface was required to allow users to test hypotheses like our own, or to undertake more exploratory and experimental visualisations of the data, as illustrated in Figure 4. The circular layout was retained, as it provided a useful method of presenting the play as a whole, while maintaining the temporal dimension of the character interactions. The character-selection menu and the scene-divisions in the outer ring were thus added in the final stage of development.

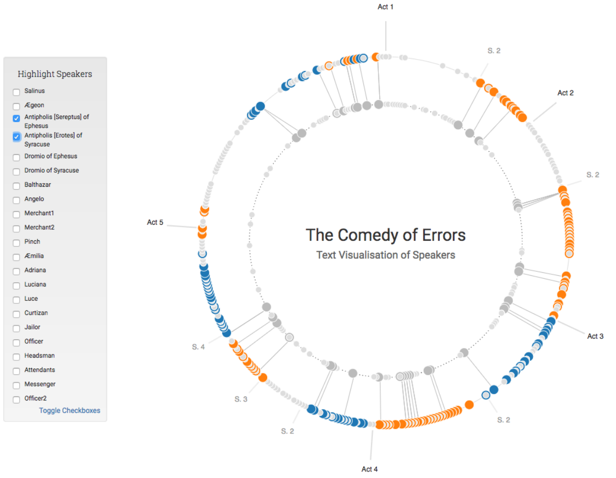

Highlight Speakers
☐ Salinus
☐ Ægeon
☑ Antipholis [Sereptus] of Ephesus
☑ Antipholis [Erotes] of Syracuse
☐ Dromio of Ephesus
☐ Dromio of Syracuse
☐ Balthazar
☐ Angelo
☐ Merchant1
☐ Merchant2
☐ Pinch
☐ Æmilia
☐ Adriana
☐ Luciana
☐ Luce
☐ Curtizan
☐ Jailor
☐ Officer
☐ Headsman
☐ Attendants
☐ Messenger
☐ Officer2
Toggle Checkboxes

Figure 4: Final interface design

Also added were visualisations of higher level metrics to illustrate the number of times a character speaks, and the number of lines they speak.

Conclusion

A major part of the tool’s value is its extensibility. It may be used to create character visualisations for any play which is XML-encoded according to quite minimal specifications, and offer the opportunity to undertake comparative analysis of structural, narrative, and character-based patterns in different plays. Indeed, while the development of the tool focused on *The Comedy of Errors,* a similar visualisation of *The Winter’s Tale* (Figure 5) was generated from New Variorum Shakespeare XML files with no revision to our code.

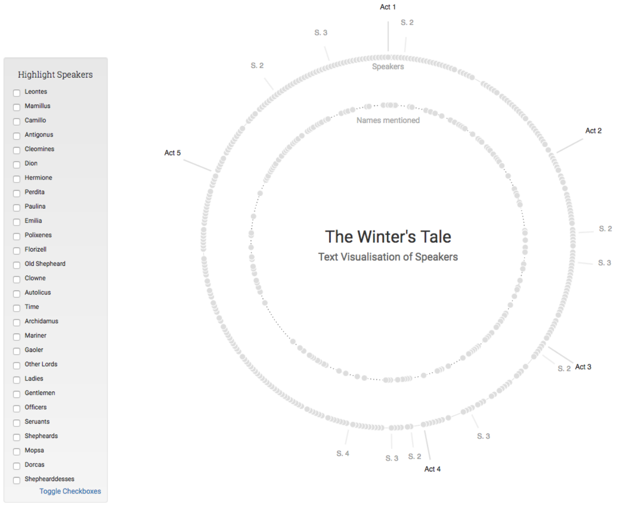

Highlight Speakers
☐ Leontes
☐ Mamillus
☐ Camillo
☐ Antigonus
☐ Cleomines
☐ Dion
☐ Hermione
☐ Perdita
☐ Paulina
☐ Emilia
☐ Polixenes
☐ Florizell
☐ Old Shepheard
☐ Clowne
☐ Autolicus
☐ Time
☐ Archidamus
☐ Mariner
☐ Gaoler
☐ Other Lords
☐ Ladies
☐ Gentlemen
☐ Officers
☐ Seruants
☐ Shepheards
☐ Mopsa
☐ Dorcas
☐ Shephearddesses
Toggle Checkboxes

Figure 5: Visualisation of The Winter’s Tale

The trajectory of *Personæ*’s development from fixity to interactivity represents a conclusion that we drew in the course of this project: that a visualisation tool developed for a particular purpose need not be confined to its use for that objective alone. The modular and open-source principles of software development have contributed to a rich and fruitful habit of sharing within the field of Digital Humanities, and we hope that others will build upon the tool that we have developed here.

Indeed, we have plans for further developments and improvements to *Personæ*. Working towards a tool which will enable structural and thematic comparison of the thirty-six plays in the First Folio, the next phase will test for structural correlations in a thematic grouping of five additional Shakespearean plays. This development will strengthen *Personæ*’s potential for generating insights into macro-level structural analysis of dramatic texts, while testing its technical extensibility by incorporating XML files from another source, [The Bodleian First Folio.]

Bibliography

Bostock, M., Ogievetsky, V., and Heer, J. (2011) "D3: Data-Driven Documents." *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011): 2301-2309. Web.

Grandjean, M. (2015) “Network Visualization: Mapping Shakespeare’s Tragedies.” Martin Grandjean. N.p., 23 Dec. Web. 26 Oct. 2016.

Moretti, F. (2011). “Network Theory, Plot Analysis.” New Left Review 68: 80–102. Print.

Stiller, J., Nettle, D., and Dunbar, R. (2003) "The Small World of Shakespeare’s Plays." *Human Nature* 14.4: 397-408. Print

Xanthos, A., Pante, I., Rochat, Y., Grandjean, M. (2016). “Visualising the Dynamics of Character Networks.” *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, Kraków, 2016. 417-419. Print.

Yose, J., Kenna, R., Carron, P. M, Platini, T., and Tonra, J. (2016). “A Networks-Science Investigation into the Epic Poems of Ossian.” *Advances in Complex Systems* (2016): 1650008. Web. 26 Oct. 2016.

A Framework for Historical Russian Flu Epidemic Exploration from German Newspapers

Tran Van Canh
ctran@l3s.de
L3S Research Center, Hannover, Germany

Katja Markert
markert@cl.uni-heidelberg.de
Heidelberg University, Germany

Wolfgang Nejdl
nejdl@l3s.de
L3S Research Center, Hannover, Germany

Introduction

The Russian flu 1889-1893 epidemic reached Europe from the East in November and December of 1889 and spread over the whole globe in the space of a few months. It was one of the first epidemics of influenza that occurred during the period of the rapid development of bacteriology. In addition, it was the first ever epidemic that was publicly and intensively narrated in the developing daily press, especially those published in German located in Germany and Austria (Mirosławska et al., 2013). However, as stated in (Valtat et al., 2011), very limited information about the epidemiology of this influenza has been found, which was based on materials published in English. While a large amount of news about the flu was published in German, it is hard to find a study on the epidemic based on German documents. These motivate our goal in this work, which is to build a framework from German materials to support research community getting more insights into the disease. Our framework consists of different components including data collection and cleaning, corpus creation, and associated tools for analysis. The framework is pictorially shown in Figure 1.

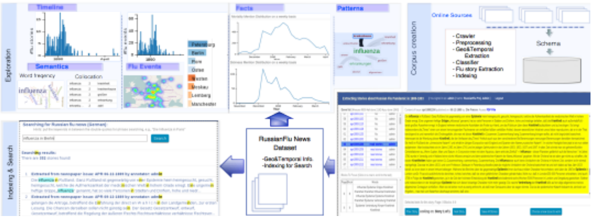


Figure 1. Russian flu exploration framework

Related work

There is limited information about the epidemiology of the Russian flu epidemic 1889-1893. In (Mirosławska et al., 2013), the authors conducted an analysis to examine the impact of the epidemic in 14 cities in Europe. Their results showed that the epidemic spread quickly from Saint Petersburg, Russia to other parts of Europe with a speed of around 400 km/week and reached the American continent only 70 days after the original peak in Saint Petersburg. In addition, some detailed information about case fatality ratio and the median basic reproduction was given also. However, their work was based on reports of only two local daily newspapers in Poznań, which implies some uncertainty due to the lack of data coverage. Valleron et al., 2010 presented a case study on the transmissibility and geographic spread of the Russian flu. A similar approach was followed by Valtat et al., 2011 to examine the age distribution of the affected people and the mortality rate of this flu event. In a recent study, Ewing et al., 2016 collected contemporary reports and explored a digital humanities approach to interpret information dissemination regarding this particular epidemic. The limitations common to all of these studies are the heterogeneity and lack of coverage of data used.

Data preparation

ID	Keyword	Variation
1	Influenza	Jnfluenza,Jnsolvenza
2	Epidemie	
3	Influenza-Epidemie	Influenzaepidemie
4	Grippe	
5	erkrankt	ertrankt
6	Pathologie	

Table 1: Keywords used to collect high recall collection of newspaper issues containing stories about influenza epidemic

Data collection

Data used in this work was collected from the Austrian Newspapers Online (ANNO) repository. ANNO contains almost all issues from many newspapers in Austria and Germany during the time the Russian flu epidemic took place. The data are accessible in both scanned PDF and OCR formats. These are appropriate for our goal in terms of extracting Russian flu related stories from noisy OCR text and checking against the scanned PDF content for validity. To establish the data collection, the keywords listed in Table 1 (along with some misspelt variations of keywords, which were included due to OCR misrecognition) were used to search the ANNO repository The search query was constrained by the time interval from 1889 to 1893. After preprocessing the search results we obtained 4,806 issues, which become the candidates to extract stories about the Russian flu.

Noise reduction

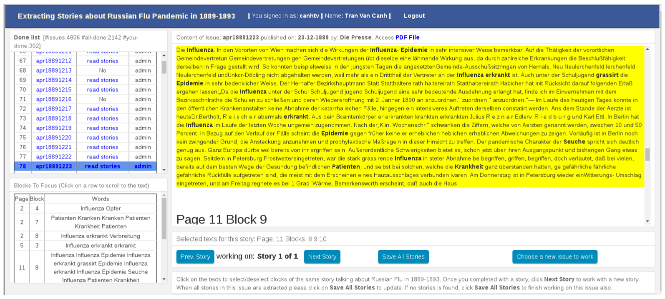
Due to the low quality of the scanned images of newspaper issues, a lot of noise is present in the corresponding OCR texts. The word error rate (WER) computed on sample texts is around 18.9%. Our goals here were to remove noise and correct misrecognized words as much as possible but at the same time manage keep the language as it was so that the derived corpus pertains its historical perspective. It is noted that modern German is rather different in writing and usage of many words due to the language's evolution. To cope with these issues, we adopted a snapshot of the Google-2-gram dataset for German from 1885 to 1895. The dataset was used to train our bigram-based model for word segmentation and spell checking. After running the model, the word error rate was reduced to 5.5%.

Text block classification

A difficult challenge for the task of extracting complete stories is that recognized OCR text blocks are very often not aligned in the same order as they were in the original image of an issue. Our approach was to automatically pre-classify OCR text blocks to identify the ones that are more likely part of some flu-related stories. Then we developed a tool to effectively help annotators extract complete Russian flu stories. For this, we adopted the KL-divergence based technique developed in (Schneider, 2004) to build a classifier. The model was trained with 245 OCR text paragraphs and obtained recall of 81.5% and precision of 68.6%. Basically, the output of the classifier can be used to help annotators start working on an issue by looking at suggested text blocks first, from which they can then select paragraphs that are part of the same story.

Extraction tool

After completing the high-recall automatic pre-extraction, we implemented a Web-based tool for annotators to help build our corpus collaboratively. The main GUI of our tool is shown in Figure 2. After having annotators work through the whole collection, we obtained a corpus of 639 news articles about Russian flu from 42 newspapers, identified with 85.7% agreement between annotators.



Extracting Stories about Russian Flu Pandemic in 1889-1893 || You signed in as: **canhtv** || Name: **Tran Van Canh** || **Logout**

Done list [#issues:4806 #all-done:2142 #you-done:302]

67	apr18891212	read stories	admin
68	apr18891213	No	admin
69	apr18891214	read stories	admin
70	apr18891215	read stories	admin
71	apr18891216	No	admin
72	apr18891217	read stories	admin
73	apr18891218	read stories	admin
74	apr18891219	read stories	admin
75	apr18891220	read stories	admin
76	apr18891221	read stories	admin
77	apr18891222	read stories	admin
78	**apr18891223**	**read stories**	**admin**

Blocks To Focus (Click on a row to scroll to the text)

Page	Block	Words
2	4	Influenza Opfer
2	7	Patienten Kranken Kranken Patienten Krankheit Patienten
2	8	Influenza erkrankt Verbreitung
5	3	Influenza erkrankt erkrankt
11	8	Influenza Influenza Epidemie Influenza erkrankt grassirt Epidemie Influenza erkrankt Influenza Epidemie Seuche Influenza Patienten Krankheit

Content of Issue: **apr18891223** published on: **23-12-1889** by: **Die Presse**. Access **PDF File**

Die **Influenza**. In den Vororten von Wien machen sich die Wirkungen der **Influenza**- **Epidemie** in sehr intensiver Weise bemerkbar. Auf die Thätigkeit der vorortlichen Gemeindevertretun Gemeindevertretungen gen Gemeindevertretungen übt dieselbe eine lähmende Wirkung aus, da durch zahlreiche Erkrankungen die Beschlußfähigkeit derselben in Frage gestellt wird. So konnten beispielsweise in den jüngsten Tagen die angesetztenGemeinde-Ausschußsitznngen von Hernals, Neu Neulerchenfeld lerchenfeld Neulerchenfeld undUnkcr-Döbling nicht abgehalten werden, weil mehr als ein Dritttheil der Vertreter an der **Influenza erkrankt** ist. Auch unter der Schuljugend **grassirt** die **Epidemie** in sehr bedenklicher Weise. Der Hernalfer Bezirkshauptmann Statt Statthaltereirath haltereirath Statthaltereirath Habicher hat mit Rücksicht darauf folgenden Erlaß ergehen lassen:„Da die **Influenza** unter der Schul Schuljugend jugend Schuljugend eine sehr bedeutende Ausdehnung erlangt hat, finde ich im Einvernehmen mit dem Bezirksschnlrathe die Schulen zu schließen und deren Wiedereröffnung mit 2. Jänner 1890 an anzuordnen." zuordnen." anzuordnen."— Im Laufe des heutigen Tages konnte in den öffentlichen Krankenanstalten keine Abnahme der katarrhalischen Fälle, hingegen ein intensiveres Auftreten derselben constatirt werden. Ans dem Stande der Aerzte ist heuteDr.Bertholt, Reischer abermals **erkrankt**. Aus dein Bcamtenkörper er erkrankten krankten erkrankten Julius Reznar Edlerv. Riedburg und Karl Ettl. In Berlin hat die **Influenza** im Laufe der letzten Woche ungemein zugenommen. Nach der„Klin. Wochenschr." schwanken die Ziffern, welche von Aerzten genannt werden, zwischen 10 und 50 Percent. In Bezug auf den Verlauf der Fälle scheint die **Epidemie** gegen früher keine er erheblichen heblichen erheblichen Abweichungen zu zeigen. Vorläufig ist in Berlin noch kein zwingender Grund, die Ansteckung anzunehmen und prophylaktische Maßregeln in dieser Hinsicht zu treffen. Der pandemische Charakter der **Seuche** spricht sich deutlich genug aus. Ganz Europa dürfte wol bereits von ihr ergriffen sein. Außerordentliche Schwierigkeiten bietet es, schon jetzt über ihren Ausgangspunkt und bisherigen Gang etwas zu sagen. Seitdem in Petersburg Frostwettereingetreten, war die stark grassirende **Influenza** in steter Abnahme be begriffen, griffen, begriffen, doch verlautet, daß bei vielen, bereits auf dem besten Wege der Gesundung befindlichen **Patienten**, und selbst bei solchen, welche die **Krankheit** ganz überstanden hatten, ge gefährliche fährliche gefährliche Rückfälle aufgetreten sind, die meist mit dem Erscheinen eines Hautausschlages verbunden ivaren. Am Donnerstag ist in Petersburg wieder einWitterungs- Umschlag eingetreten, und am Freitag regnete es bei 1 Grad 'Wärme. Bemerkenswcrth erscheint, daß auch die Haus

Page 11 Block 9

Selected texts for this story: Page: 11 Blocks: 8 9 10

Click on the texts to select/deselect blocks of the same story talking about Russian Flu in 1889-1893. Once you completed with a story, click **Next Story** to work with a new story. When all stories in this issue are extracted please click on **Save All Stories** to update. If no stories is found, click **Save All Stories** to finish working on this issue also.

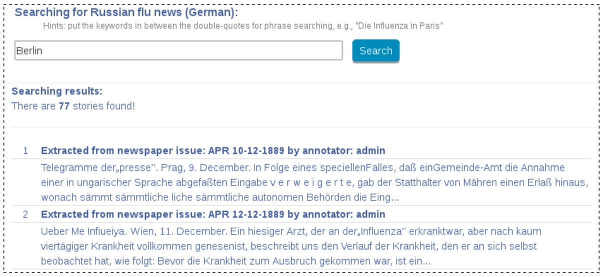
Figure 2. Main GUI of our tool for Russian flu story extraction

Geo and temporal information extraction

Given that location and time are helpful features for exploring the development of the epidemic, we extracted and normalized geographic names and temporal expressions occurring in the corpus. For geographic names, the Geodict tool created by Pete Warden (2011) was adapted to work with country and city names in German. HeidelTime (Strötgen and Gertz, 2013) was used for temporal information extraction and normalization.

Indexing and search engine

We created a search engine on the corpus to support research community in searching for information. The searching GUI is shown in Figure 3.



Searching for Russian flu news (German):

Hints: put the keywords in between the double-quotes for phrase searching, e.g., "Die Influenza in Paris"

Searching results:

There are **77** stories found!

1 **Extracted from newspaper issue: APR 10-12-1889 by annotator: admin**

Telegramme der„presse". Prag, 9. December. In Folge eines speciellenFalles, daß einGemeinde-Amt die Annahme einer in ungarischer Sprache abgefaßten Eingabe verweigerte, gab der Statthalter von Mähren einen Erlaß hinaus, wonach sämmt sämmtliche liche sämmtliche autonomen Behörden die Eing...

2 **Extracted from newspaper issue: APR 12-12-1889 by annotator: admin**

Ueber Me Infiueiya. Wien, 11. December. Ein hiesiger Arzt, der an der„Influenza" erkranktwar, aber nach kaum viertägiger Krankheit vollkommen genesenist, beschreibt uns den Verlauf der Krankheit, den er an sich selbst beobachtet hat, wie folgt: Bevor die Krankheit zum Ausbruch gekommen war, ist ein...

Figure 3. Russian flu story searching module

Exploration tools and sample results

We provided associated tools along with the corpus. The corpus timeline provides statistics on the number of stories in the corpus across time and news outlet. In addition, it provides interactive visualization. As an example shown in Figure 4, during the peak time in late December 1889 and January 1890, extensive news about the influenza was published.

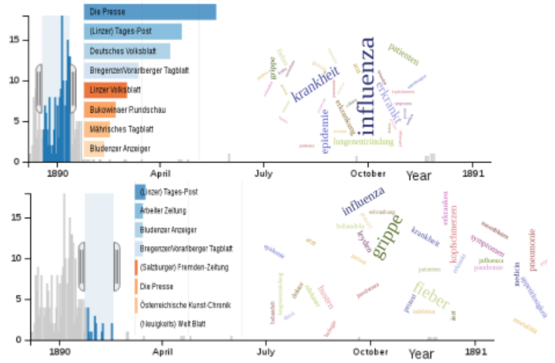


Figure 4. Press attention on the flu and topic changes over time

Newspapers were trying to narrate the outbreak as fast as possible. Words that appear significantly in the stories include *influenza*, *epidemic*, *krankheit (disease)* and *erkrankt (sick)*. A short time after this peak period, fewer reports were published about the outbreak of the flu and communities started discussing the treatment more. Names of doctors appearing in the news (e.g., Leyden, Proust) together with words describing symptoms such as *fieber (fever)*, *kopfschmerzen (headache)*, *appetitlosigkeit (anorexia)*. Furthermore, by exploring word collocations one can find even more interesting information. Figure 5 shows a frequent pattern of word collocation describing the influenza. The words *heute (today)* and *gestern (yesterday)* indicate that news about the flu is updated every day*; and* the word *jänner (January)* implies that the flu outbreaks happened during winter.

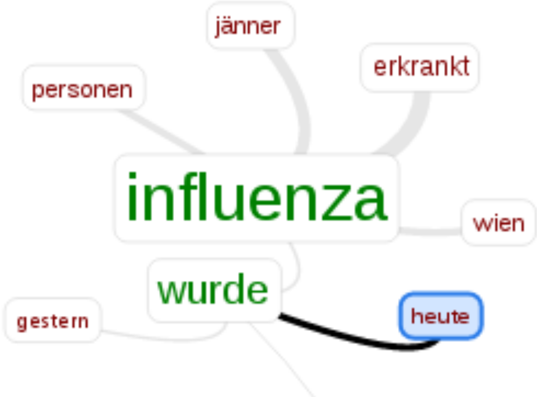


Figure 5: A frequent pattern of word collocation extracted from our corpus

Figure 6 presents the co-occurrences of three words *infuenza, erkrankt, and london* over time. It shows a similar trending pattern of the words *infuenza* and *erkrankt* being used to note about the flu. In addition, one can observe that the peak time of the flu in London was from late December 1889 to early January 1890 as indicated in (Honigsbaum, 2010; Goff, 2011). This suggests that the temporal distribution of terms can give us more insights into the evolution of the epidemic.

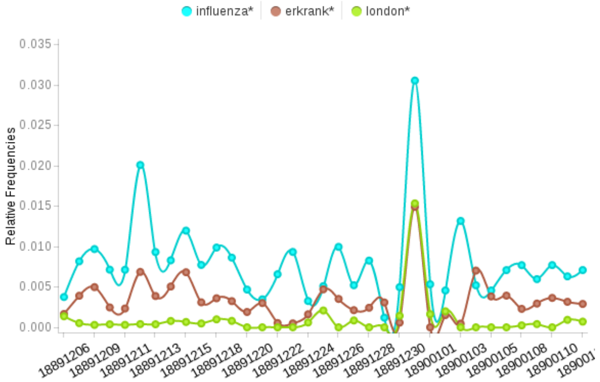


Figure 6: Similar trending pattern of the two words influenza and erkrankt, and the peak time of the flu in London

The framework also supports studying the evolution of the flu over time and geographic regions. We employed the method introduced in (Abdelhaq et al., 2013) for localized event detection. Figure 7 shows three snapshots describing the development of the epidemic over cities in Europe during the peak time from late November 1889 to January 1890.

Summary

We have introduced a framework for research communities to explore the historical Russian flu 1889-1893 from German newspapers. We developed a tool for collaborative annotators to help build our corpus. We further presented some interesting insights that we achieved from analyzing articles in the corpus. By making the corpus and associated tools available, we provide useful contributions to the community in support of conducting studies on influenza epidemics and evaluating temporal IR models.

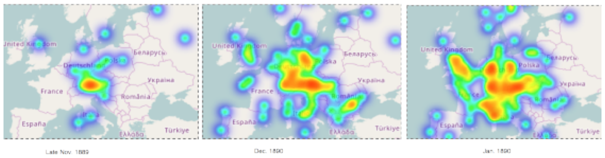


Figure 7. Evolution of the Russian flu over geographic regions during its peak time

Acknowledgements

This research is supported by the German Research Foundation (DFG) for the project "Tracking the Russian Flu in U.S. and German Medical and Popular Reports, 1889-1893" on Grant No. NE 638/13-1. We also thank you the Austrian National Library for supporting us in collecting data.

Bibliography

Abdelhaq, H., Sengstock, C., Gertz, M..(2013). "EvenTweet: Online Localized Event Detection from Twitter." Proc. VLDB Endowment Journal, 6(12):1326-4.

Aramaki, F., Maskawa, S., Morita, M..(2011). "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter." In proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1568-9.

Austrian National Library. (2011) Austrian Newspapers Online. Repository online at http://anno.onb.ac.at

Ewing, E.T., Kimmerly, V. and Ewing-Nelson, S. (2016). "Look Out for 'La Grippe': Using Digital Humanities Tools to Interpret Information Dissemination during the Russian Flu, 1889—90." Medical history,60(1):129-3.

Honigsbaum, M.(2010). "The Great Dread: Cultural and Psychological Impacts and Responses to the Russian Influenza in the United Kingdom 1889–1893".Social history of medicine,23: 299-21.

Kempińska-Mirosławska, B., and WoŸniak-Kosek, A.(2013). "The influenza epidemic of 1889–90 in selected European cities – a picture based on the reports of two Poznań daily newspapers from the second half of the nineteenth century." Med Science Monitor,19:1131-11.

Le Goff, J.M.(2011). "Diffusion of influenza during the winter of 1889-1890 in Switzerland." Jenus, 67(2): 77-23.

Paul, M.J. and Dredze, M.. (2011). "You Are What You Tweet: Analyzing Twitter for Public Health." In proceedings of the Fifth International Conference on Weblogs and Social Media. pp. 265-8.

Schneider, K.-M.(2004). "A New Feature Selection Score for Multinomial Naive Bayes Text Classification Based on KL-divergence." Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions. pp. 186-4.

Strötgen, J. and Gertz, M..(2013). "Multilingual and cross-domain temporal tagging." Language Resources and Evaluation, 47(2): 269-30.

Valleron, A.J., Cori, A., Valtat, S., Meurisse, S., Fabrice Carrat, F., and Boëlle, P.Y.. (2010). "Transmissibility and geographic spread of the 1889 influenza pandemic." In proceedings of the National Academy of Sciences of the United States of America (PNAS). pp. 8778-4.

Valtat, S., Cori, A., Carrat, F., and Valleron, A.-J. (2011). "Age distribution of cases and deaths during the 1889 influenza pandemic."Vaccine, 29(2): B6-B10.

Warden, P. (2010) GeoDict. Accessible via Github at https://github.com/petewarden/geodict

Annotonia: Annotations from Browser to TEI

Greg Tunink
techgique+conftool@unl.edu
University of Nebraska-Lincoln Libraries
United States of America

Karin Dalziel
kdalziel2@unl.edu
University of Nebraska-Lincoln Libraries
United States of America

Jessica Dussault
jdussault@unl.edu
University of Nebraska-Lincoln Libraries
United States of America

Emily J. Rau
erau2@unl.edu
University of Nebraska-Lincoln Libraries
United States of America

Introduction

The *Willa Cather Archive (WCA)* at the University of Nebraska-Lincoln (UNL) is currently working on transcription and annotation of 1500 letters to be released in 2018. As editors will write several thousand annotations, the workflow logistics are complicated. Annotonia (a portmanteau of "annotation" and *My Ántonia*, a 1918 Willa Cather novel) is a solution developed within the Center for Digital Research in Humanities (CDRH) that allows editors to write annotations directly on letters in a browser and insert those annotations into Text Encoding Initiative (TEI) XML files. Multiple editors review annotations, track letters' annotation statuses, and generate a new TEI file incorporating the annotations, avoiding having to manually edit each file. Annotonia utilizes both pre-existing, customized open source software and new software developed for this project. This paper describes the difficulties faced, the workflow of Annotonia, and its prospects for future annotation work.

The challenge

The *Complete Letters of Willa Cather* is a National Endowment for the Humanities-funded project to publish a digital, fully annotated edition of the letters of Willa Cather, a 20th century American novelist. The project includes student editorial assistants, staff, and faculty both at UNL and working remotely. Because of differences in locale, technical abilities, and Cather-related expertise, a solution was sought for drafting, revising, and encoding letter-specific annotations that would fit the skillsets of all collaborators.

The *WCA* has two types of annotations: authority files and letter-specific information. Editors skilled with TEI XML curate and encode people, places, and works shared across the corpus into authority files. Letter-specific annotations are more challenging to manage as they are written by a wider group of scholars, many unfamiliar with encoding XML. Editors previously used a cumbersome process of pasting documents into Word files, sharing them for annotation, and laboriously copying received annotations back into the XML. This tedious process introduced errors and was deemed unsustainable for the large number of anticipated letter-specific annotations.

With these difficulties in mind, *WCA* editors collaborated with CDRH developers to envision tools that would allow viewing the content of letters with all associated materials; writing letter-specific annotations that might include images, links, or other materials; and exporting finalized versions of these annotations as TEI XML. Editors also identified the need to view annotations that had already been written so information contained in them would not be repeated.

Project requirements

- Display letters via HTML with existing controlled vocabulary annotations for people, places, and works.
- Provide an interface where editors can browse, edit, approve, and flag annotations.
- Export annotations to their preservation format, TEI P5 XML.
- Insert <ref> tags into the TEI XML corresponding to ID's in the annotation file above.

Existing software review

Before building Annotonia, we looked at existing technologies and methods for annotating HTML. Web-based annotation is not a new concept, by any means. The value of annotating HTML has long been recognized for its collaborative strengths, allowing users to identify errors, review, comment on, and bookmark documents (W3C Digital Publishing Interest Group, 2014). The W3C currently has a Web Annotation Working Group dedicated to creating specifications for "interoperable, sharable, distributed Web annotation architecture" (W3C Web Annotation Working Group, 2017; Web Annotation Data Model, 2017).

Most annotation software reviewed was not suitable for the project. Many existing solutions (such as Annotation Studio, Recogito2, and Editor's Notes) did not allow in place annotation, instead requiring one to upload documents to their system. This would strip out the existing annotations, making work more difficult for editors. Some (such as PundIT) were not fully open source. Others (Such as Hypothes.is) were developed for public input and would require stripping out many features for our needs. Other problems encountered included installation roadblocks, lack of maintenance, and poor documentation. Importantly, our requirement of exporting annotations and embedding them into TEI documents was not supported by any existing systems.

With these considerations, programmers chose to work with Annotator and its back end complement, Annotator Store, because the software interacts through an API and uses XPaths for pinpointing annotations' locations (Open Knowledge Foundation, 2016a; 2016b). Annotator proved to be easy to install and extend with community plugins, such as rich text editing, keyword tagging, and filtering. Annotator Store's use of Elasticsearch reinforced the decision to use it, as Elasticsearch is widely used and receives generous community development and support. Annotator deals only with the annotation part of the requirement and not categorizing or workflow, which had to be developed in-house.

Annotonia: the solution

The first step was simply to display the letters' TEI XML in a browser in a way that minimally rearranged the structure of the documents. TEI Boilerplate uses browser XSLT capabilities to create an HTML/TEI hybrid representation of TEI documents with small alterations for links, images, and other elements (Walsh et al. 2013). Boilerplate was therefore useful for constructing HTML from documents, while allowing for TEI files to be added and removed quickly from the site structure and workflow.

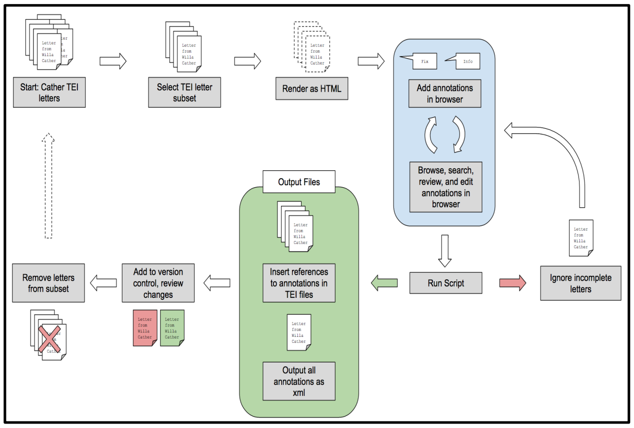
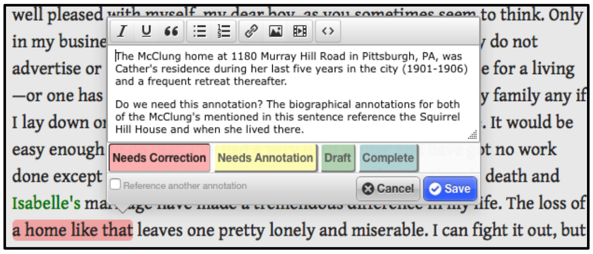

Start: Cather TEI letters
Select TEI letter subset
Render as HTML
Fix
Info
Add annotations in browser
Browse, search, review, and edit annotations in browser
Run Script
Ignore incomplete letters
Output Files
Insert references to annotations in TEI files
Output all annotations as xml
Add to version control, review changes
Remove letters from subset
Letter from Willa Cather

Figure 1: The Annotonia workflow for a single batch of letters

In order to be able to annotate the documents displayed using TEI Boilerplate, programmers embedded Annotator in the web page and modified it so it had suitable options for the *WCA* editors. These included stripping out some of the text editing capabilities and adding workflow-specific options.

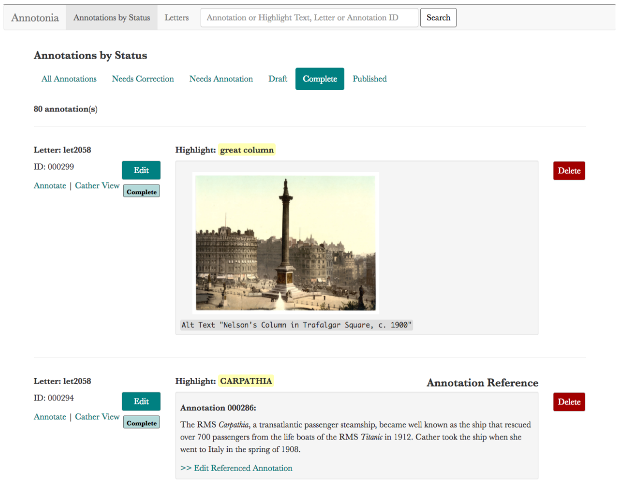

well pleased with myself, my dear boy, as you sometimes seem to think. Only
in my busine ... y do not
advertise or ... e for a living
—or one has ... y family any if
I lay down o ... . It would be
easy enough ... t no work
done except ... death and
Isabelle's ma ... life. The loss of
a home like that leaves one pretty lonely and miserable. I can fight it out, but

The McClung home at 1180 Murray Hill Road in Pittsburgh, PA, was Cather's residence during her last five years in the city (1901-1906) and a frequent retreat thereafter.

Do we need this annotation? The biographical annotations for both of the McClung's mentioned in this sentence reference the Squirrel Hill House and when she lived there.

Figure 2: An example of creating and categorizing an annotation with Annotonia

Annotator does not include an interface for searching, browsing, and editing all of the annotations from the Annotator Store. PHP pages were written to provide these capabilities.



Annotations by Status

80 annotation(s)

Letter: let2058
ID: 000299

Highlight: great column

Alt Text "Nelson's Column in Trafalgar Square, c. 1900"

Letter: let2058
ID: 000294

Highlight: CARPATHIA

Annotation 000286:

The RMS *Carpathia*, a transatlantic passenger steamship, became well known as the ship that rescued over 700 passengers from the life boats of the RMS *Titanic* in 1912. Cather took the ship when she went to Italy in the spring of 1908.

Figure 3: Annotation review page including an annotation with an image and a referenced annotation

The last requirement for Annotonia was a conversion script that inserts annotation references into TEI documents while exporting an authority file containing the annotations. The conversion script takes a subset of XML files, queries the Annotator Store API, adds tags for annotations based on XPaths, and outputs a list of possibly incorrect insertions that require review.

Use and extension

The collection of new scripts and open source software comprising Annotonia has been able to handle the workflow requirements of the *WCA's* letter annotations, though there is room for improvement. The areas that require the most attention are the rendering of TEI in HTML and the scripted editing of TEI files. TEI Boilerplate necessarily alters the TEI in order to mimic the behavior of HTML links and images, which means that occasionally the location of an annotation in the HTML view is difficult to match up programmatically due to differing XPaths.

A new alternative to Boilerplate, CETEIcean, which promises to preserve "the full structure and information from your TEI data model," may be one possibility to address this problem (Cayless and Viglianti, 2016). Programmers would need to evaluate how easily it can be incorporated into the Annotonia workflow, as well as the ability to load the Annotator JavaScript libraries in the page.

Some aspects of the Annotonia code base are tailored for use only with TEI Boilerplate display and *WCA* file naming conventions. These functions would have to be generalized to make Annotonia easier to integrate with other tools and projects. Meanwhile, the Annotator community continues to improve the software. A new version is forthcoming which includes modifications to enhance the experience of creating, saving, and updating the highlighting of HTML content. When this version of Annotator is released, it may require some reworking of existing customizations, but updating Annotonia to incorporate its new features will support a broader variety of projects.

Reception

The *WCA* began using Annotonia in September 2016. Guidelines and instructional videos demonstrating Annotonia's functions have alleviated difficulties from the varying technical skills of editors. Although still in the beginning stages of annotating individual letters, the tool works well for collaborating on, drafting, and revising annotations. Editors have estimated that each annotation automatically handled saves around 5 minutes of time, so the potential time savings is several months of work. Through an iterative process, improvements have been added as users uncover inefficiencies and provide feedback. By 2018, several thousand annotations will be created with Annotonia and published with the complete letters on the *WCA*.

Further applications

A frequent difficulty for digital archival projects is efficiently proofreading and annotating texts. In the CDRH, these workflows on other projects resemble the *WCA*'s old process of marking up Word documents, or, worse yet, printing out entire websites and annotating by hand. Designing solutions by combining open source software with well documented, configurable scripts and workflows has proven to be effective in providing flexibility to cover a variety of needs. As we apply this method to extending and generalizing Annotonia for other projects like the *Walt Whitman Archive* and *The William F. Cody Archive*, we will further refine deployment and documentation.

The PHP and conversion components of Annotonia have been published ("Annotonia Status" and "Annotonia Converter", and the customized pieces of existing software will be published soon. Full publication of Annotonia will involve further documentation and testing. It is the hope of the Annotonia team that this tool will not only prove to be useful internally, but will provide inspiration to other text based editions seeking to automate annotation processes.

Bibliography

Cayless, H. A. and Viglianti, R. (2016). "TEIC/CETEIcean." *GitHub*. https://github.com/TEIC/CETEIcean (accessed 19 October 2016)

Open Knowledge Foundation (2016a). "Annotator - Annotating the Web." http://annotatorjs.org/ (accessed 11 October 2016).

Open Knowledge Foundation (2016b). "Annotator-Store." *GitHub*. https://github.com/openannotation/annotator-store (accessed 11 October 2016).

W3C Digital Publishing Interest Group and Open Annotation Community Group (2014). "Annotation Use Cases." http://www.openannotation.org/usecases.html (accessed March 3, 2014).

W3C Web Annotation Working Group (2017). "Web Annotation Data Model." https://www.w3.org/TR/annotation-model/ (accessed February 23, 2017).

W3C Web Annotation Working Group (2016). "W3C Web Annotation Working Group." https://www.w3.org/annotation/ (accessed October 11, 2016).

Walsh, J., Simpson, G., and Moaddeli, S. (2016) "TEI Boilerplate." http://dcl.ils.indiana.edu/teibp/ (accessed 19 October 2016).

Willa Cather Archive. "Github Organization." *GitHub*. https://github.com/Willa-Cather-Archive (accessed 19 October 2016).

Verifying the Authorship of Saikaku Ihara's Arashi ha Mujyō Monogatari in Early Modern Japanese Literature: A Quantitative Approach

Ayaka Uesaka
auesaka@mail.doshisha.ac.jp
Organization for Research Initiatives and Development
Doshisha University, Japan

Introduction

This study focuses on *Arashi ha Mujyō Monogatari ("The Tale of Transient Popular Kabuki Actor Arashi's Life"; 1688)*, a novel from the early modern Japanese literature, written by Saikaku Ihara (1642–93). It is a first work of a Kabuki actor's life in Japan (Kabuki is a traditional stage arts performed exclusively by male actors with the accompaniment of live music and songs). Then we will examine the "authorship problem" in Saikaku's works using the tools of quantitative analysis.

Saikaku was a national author whose novels were published in 17th century. Saikaku's works are known for their significance for developing Japanese contemporary novels. One recent hypothesis has stated that he wrote twenty-four novels, however, it remained unclear which works were really written by Saikaku except *Kōshoku ichidai otoko ("The Life of an Amorous Man"; 1682)*, *Shōen Ōkagami ("The Great*

Mirror of Female Beauty"; 1684), Kōshoku ichidai onna ("The Life of an Amorous Woman"; 1686), Kōshoku gonin onna ("Love Stories about Five Women"; 1686). Although the study of his works has continued, these fundamental doubts about his authorship remain.

Meanwhile, the potential of quantitative analysis of textual data and the related field of the digital humanities have also dramatically advanced. However, quantitative analysis of Japanese classical works has been behind. It has been a problem due to complications regarding development of morphological analysis software and also delayed digitalization of Japanese classical works.

Previous Studies

Found by Noma in 1941

Noma found and introduced *Arashi ha Mujyō Monogatari* in 1941. He mentioned that the novel was actually written by Saikaku, for the following reasons (Noma, 1941 and 1964). (1) The handwriting of the novel belongs to Saikaku; and (2) He found a similar writing error in *Arashi ha Mujyō Monogatari* and Saikaku's work.

Arguments for Saikaku's authorship

The handwriting is not crucial in deciding if they are Saikaku's novels. According to Emoto *et al.* (1996), among his twenty-four novels, the handwriting of nineteen works does not belong to Saikaku. Moreover, Saikaku made a fair copy of other writer's draft such as *Kindai Yasa Inja ("The story of a hermit"; 1686)* by Kyōsen Sairoken (? -?) and *Shin Yoshiwara Tsurezure ("The book of commentary on the licensed quarters of a certain area"; 1689)* by Sutewaka Isogai (? -?).Mori (1955) has argued that Saikaku's novels are an apocryphal work mainly written by Dansui Hōjō (1663-1711) except *Kōshoku ichidai otoko.*

As he gained a national audience, Saikaku was pressured to write on demand and in great volume. At first he wrote only one or two novels a year, however in the two years from 1687 to 1688 he published twelve books, with a total of sixty-two volumes. Saikaku's style and approach also changed at this point (Shirane, 2004).

It is possible that Saikaku had some assistance (Nakamura,1969). *Arashi ha Mujyō Monogatari* was published in this period. Moreover, *Arashi ha Mujyō Monogatari* does not have a preface, epilogue, signature, namely it is not specified that it was written by Saikaku. Despite the authorship problem of *Arashi ha Mujyō Monogatari* remains unanswered, little work has been done about it. For that reason, this study re-examines the authorship of *Arashi ha Mujyō Monogatari* using a quantitative approach.

Databases

Database of Saikaku's Works

First, we digitized all the text of 120 works of Saikaku (24 novels, 80 poem books, etc.) based on the first edition of each works (see Figure 1). Second, Since Japanese sentences are not separated by spaces, we built the rule with early modern Japanese researchers, who were editors of *Shinpen Saikaku Zenshu ("The new complete works of Saikaku")*. Finally, based on this rule, we added spaces between the words in all of the sentences. In addition, the grammatical categories' information was added. According to our database, there are 710,355 words contained in his 120 works.

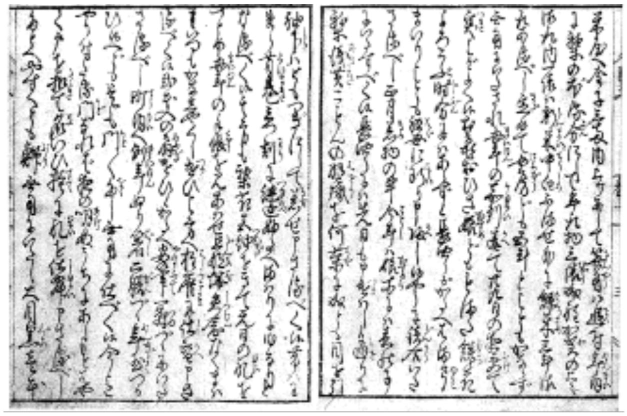


Figure 1: Saikaku's publication

Database of Dansui's Works

We also made the database of Dansui's novels Shikidō Ōtuzumi ("The Great Drum of Love"; 1687), Chuya yōjin ki ("The Night and Day of Precaution"; 1707) and Budō hariai Ōkagami ("The Great Mirror of Martial Arts"; 1709), using same methods and rules of Saikaku's database. According to our database, there are 53,838 words contained in these works. The next section considers the doubts about the authorship problem of Arashi ha Mujyō Monogatari.

Analysis and Results

In our previous studies, we have analyzed Saikaku and Dansui's novels, and have clarified the following two points by extracting their writing style using principal component analysis (PCA) and cluster analysis (hierarchical clustering): (1) A comparison of the Saikaku and Dansui's novels showed ten prominent features: the grammatical categories, words, nouns, particles, verbs, adjectives, adverbs, adnominal adjectives, grammatical categories bigrams and particle bigrams (Uesaka, 2015, 2016); and (2) Using these features, we analyzed Saikaku's four posthumous novels (many researchers have raised questions about the authorship, because these novels were edited and published by Dansui after Saikaku's death). We found these four posthumous works indicated same features of Saikaku's novel, therefore we concluded that these four posthumous novels belonged to Saikaku (Uesaka・Murakami,2015ab, Uesaka, 2016).

In this study, we compared *Arashi ha Mujyō Monogatari* to Saikaku and Dansui, as authenticated novels of them (see Table 1) by ten prominent features using PCA and cluster analysis to see the differences in each novels. The analysis revealed differences of writing style between *Arashi ha Mujyō Monogatari*, Saikaku and Dansui.

Saikaku's novels	*Kōshoku ichidai otoko, Shōen Ōkagami, Kōshoku ichidai otnna* and *Kōshoku gonin onna*
Dansui's novels	*Shikidou otsuzumi, Chuya youjin ki and Budou hariai okagami*

Table 1: The authenticated novels of Saikaku and Dansui

We conducted PCA with correlation matrix and these novels fall into three groups: Saikaku, Dansui and *Arashi ha Mujyō Monogatari* (see Figure 2). Furthermore, we conducted a cluster analysis. There also appears to be a considerable difference among *Arashi ha Mujyō Monogatari*, Saikaku and Dansui's novels. When calculating distances between each novels, we normalized the frequency of each words, and used the Kullback–Leibler divergence and the algorithm from the Ward method. Furthermore, we obtained similar result of the other nine features: the grammatical categories, words, nouns, particles, verbs, adjectives, adverbs, adnominal adjectives and particle bigrams.

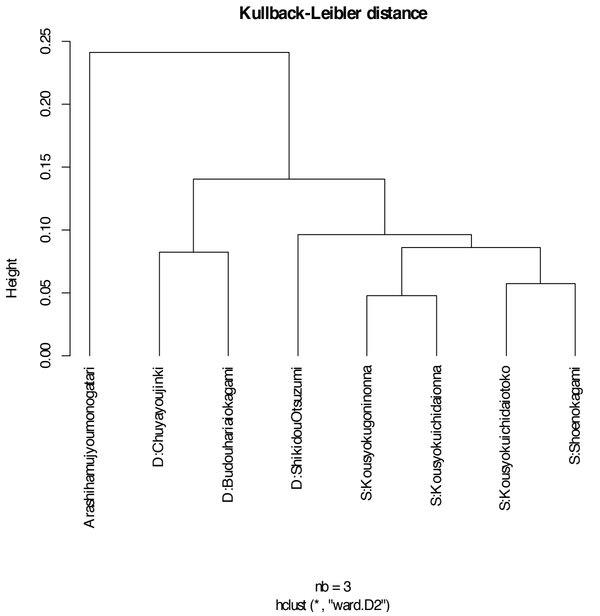


Figure 2: PCA results for grammatical categories bigrams

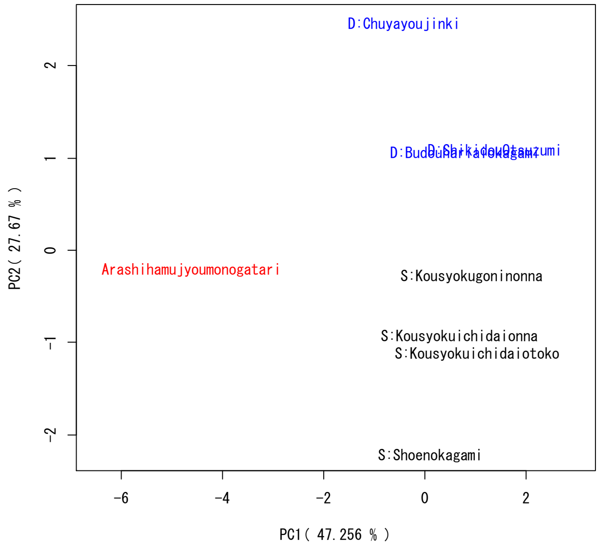


Figure 3: Cluster analysis results for grammatical categories bigrams

Discussion and Conclusion

When comparing ten prominent features using PCA and cluster analysis, we found that *Arashi ha Mujyō Monogatari* was significantly different from Saikaku and Dansui's works. A number of features indicate that *Arashi ha Mujyō Monogatari* is not Saikaku and Dansui. In order to clarify *Arashi ha Mujyō Monogatari*'s author, we need to conduct more detailed analysis. It is necessary to add the data of other writers with the possibility of the author of *Arashi ha Mujyō Monogatari*, for example, Kiseki Ejima(1666-1735) and Ichirōemon Nishimura(?-1969). We have been building the database of these author's 13 novels, and we will do comparisons in the future study.

Acknowledgements

We would like to thank Professor Masakatsu Murakami, Professor Hidekazu Banno and Professor Takayuki Mizutani for their help on our research. This work was mainly supported by the Japanese Society for the Promotion of Science (JSPS) Grant-in-Aid for Research Activity Start-up(16H07328).

Bibliography

Emoto, Y. and Taniwaki, M. (1996). Saikaku Jiten ("A Saikaku Dictionary"). Tokyo:Ouhu Publishing.

Mori, S. (1955). Saikaku to Saikaku Bon ("Saikaku and Saikaku's Novel"). Tokyo:Motomotosha Publishing.

Nakamura, Y. (1969).Saikaku Nyumon("The Introduction of Saikaku's Research"). In:Kokubungaku kaishaku to kansho(" Japanese literature-Explanation and Appreciation") 34(11). pp.10-25. Tokyo: Shibundo Publishing.

Noma,K. (1941). Arashi ha Mujyō Monogatari ("The Tale of Transient Popular Kabuki Actor Arashi's Life-Explanation and Understaning"). In: Saikaku Shin Shinkō ("Saikaku New Article";1981). pp231-290. Tokyo:Iwanami Publishing.

Noma,K. (1964). Sairon Arashi ha Mujyō Monogatari ("Re-explanation of the Tale of Transient Popular Kabuki Actor Arashi's

Life"). In: Saikaku Shin Shinkō ("Saikaku New Article";1981). pp291-313. Tokyo:Iwanami Publishing.

Shirane, H. (2004). Early Modern Japanese Literature: An Anthology 1600–1900. New York: Columbia University Press.

Uesaka, A. (2015). A Quantitative Comparative Analysis for Saikaku and Dansui's Works, Proceedings of Japan-China Symposium on Theory and Application of Data Science, pp.41〜46, Kyoto:Doshisha University Faculty of Culture and Information Science.

Uesaka, A. & Murakami, M. (2015a). Verifying the Authorship of Saikaku Ihara's Work in Early Modern Japanese Literature: A Quantitative Approach. Digital Scholarship in the Humanities. 30(4). pp.599〜607. Oxford: Oxford University Press.

Uesaka, A.& Murakami, M. (2015b). A Quantitative Analysis for the Authorship of Saikaku's Posthumous Works Compared with Dansui's Works. Proceedings of the Digital Humanites2015.

Uesaka, A. (2016). Saikaku Ikōshu no Chosha no kentō ("Verifying the Authorship of Saikaku's Posthumous Works"). pp187-263. In: The Computational Authorship Attribution. Tokyo: Bensei Publishing.

LIMORIES: Expanding Access to Local Histories and Memories with Computational Aids in the Indian Context

Lakshmi Valsalakumari
lakshmi.valsala@research.iiit.ac.in
Centre for Exact Humanities, IIIT-Hyderabad, India

Introduction

> "It is memory that makes your identity. If your memory be lost, how will you be the same man?"
> – Voltaire

India is well-known for its diversity of culture. Like its languages, its localities too have evolved through the ages, each abounding in markers that talk of its past - externalized, at times, in monuments or historical artefacts, but captured even more eloquently, perhaps, in the customs, rituals, traditions, songs, arts and crafts, language and the daily life of its people.

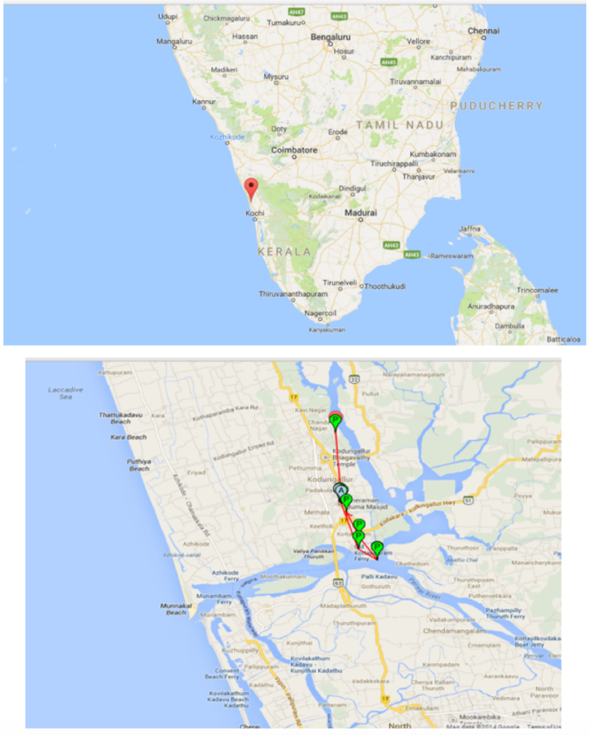
Yet, as India modernizes, this very diversity that has been a hallmark of the civilization is under threat. The homogenizing effects of globalization are all too visible as one traverses the country today; as cities and towns, that look more and more alike, expand to swallow up rural suburbs, and people give up traditional vocations and lifestyles to join the global workforce (Varghese, 2013). While this is reported positively in economic circles, and is counted as a welcoming sign of the increasing prosperity of vast majorities of the Indian populace, an unfortunate side-effect, concurrently felt, is the loss of the sense of uniqueness, traditionally associated with each Indian locality. This has led to growing pangs, across significant cross-sections of the society, about a gnawing loss of cultural identity – a concern not restricted to India or Indians (Khair, 2016).

Limories is a work-in-progress system that attempts to capture the distinctiveness and uniqueness of a locality from its people. Limories believes that the sense of past of a locality, unique to each locality, inheres in each and every individual associated with it, and is special and key to understanding and appreciating that locality wholly and in full. Limories attempts to do this by sourcing the externalized memories from the people themselves; collating, curating and presenting them - using open-source technologies. Goals include coming up with a curation framework, tools, standards and best practices, practical and relevant to the Indian context, which can be applied across localities. The generalization is a particular challenge, given the emphasis on the 'vaiseshik' – the 'distinctive' of each locality, aimed to be captured. And indeed, how one applies the general to the particular, remains a key differentiator between sciences and humanities (Singh, 2003).

This short paper will describe the pilot site where Limories was attempted, and the results thereof. It will also outline challenges being worked out, and the overall roadmap.

The pilot site

Kodungallur is a quiet town in Kerala, situated near the mouth of where one of the two arms of the perennial Periyar River empties itself into the Arabian Sea.



Source: Google Maps

Known during the colonial period as Cranganore, Kodungallur was earlier famous as capital to the Perumals of Kerala during the formative 9th-12th centuries CE of Kerala history and culture (Narayanan, 1972).

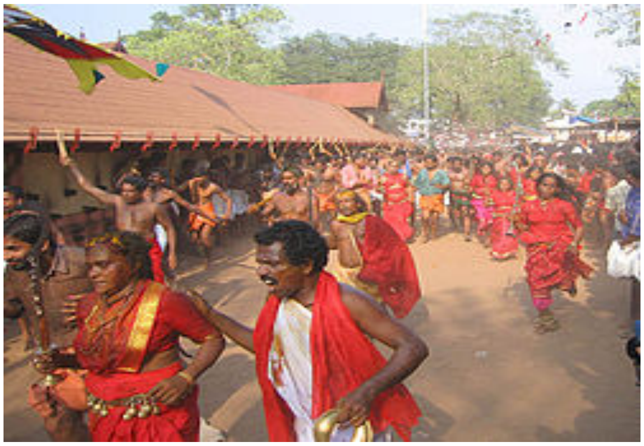
Until recently, Kodungallur was considered to be the ancient port of Muziris, described in Greco-Roman texts, such as the 'Periplus of the Erythrean Sea' (Casson, 1989), and referred to as Muchiri, in Tamil Sangam poetry. While recent research points towards Pattanam, a place about 9 km upstream from Kodungallur, as a probable contender for the exact location of the ancient port (Cherian et al., 2014), even today, the reputation of Kodungallur as the entry point of multiple religions – Christianity, Judaism and Islam, into the Indian sub-continent from across the seas, is proudly cherished by the people of the locality.

The Cheraman Perumal Mosque in Kodungallur, probably the earliest in India, its architecture providing testimony to the region's syncretic past (Picture: Own)

Among the various temples that dot the town, the one that attracts the most devotees today, almost eponymous with the town, is the Kodungallur Amma temple – dedicated to the Mother Goddess – the Mother of Kodungallur.

The Kodungallur Bhagavathy Temple or Sree Kurumba Kaavu (Picture: Own)

The temple, or 'kaavu' (sacred grove), has, in public perception, a kind of notoriety associated with its annual 'Bharani' festival – gained from the festival's rather peculiar rituals. During the Bharani period, devotees in large numbers throng to the temple from distant areas, clad in vivid red, clanging sickles in their hands, singing profane songs. The culminating 'Kaavu theendal' ritual of the festival is widely commented on and studied.



The Kaavu Theendal (Source: Wikipedia)

It was this site that the noted local historian, late Dr.N.M. Nampoothiri, former Dean of the Centre for Heritage Studies, Kerala, suggested to this researcher as a pilot. Over the years of 2014/2015, as I started gathering the history and lore related to Kodungallur and the temple from various sources, it was the Bharani rituals that stood out and piqued one's curiosity the most - with its peculiarity, and yet, its unquestionable popularity that drew faithful crowds.

'Who' were the people who came for this festival? What drew them? As I went through the articles written about the festival, I saw that theorizations ranged from it having been a practice instituted to drive out Buddhist monks, to the shrine being the memorial of the Silappathikaram protagonist Kannaki. But none of those served to explain the relevance of the festival today - why it continued to thrive and attract people. What also struck one was the contrast between the general perceptions about the festival, and the attitude of those intimately associated with the locality and the festival.

In Limories, I have presented the Kodungallur Bharani, as I saw and experienced it, during 2014 and 2015. My emphasis has not so much been on the past, as on the present, as I tried to capture what had fascinated me - the continuities and connections the locale and the festival had nurtured and preserved, across time and place.

The presentation – as a photo–narrative

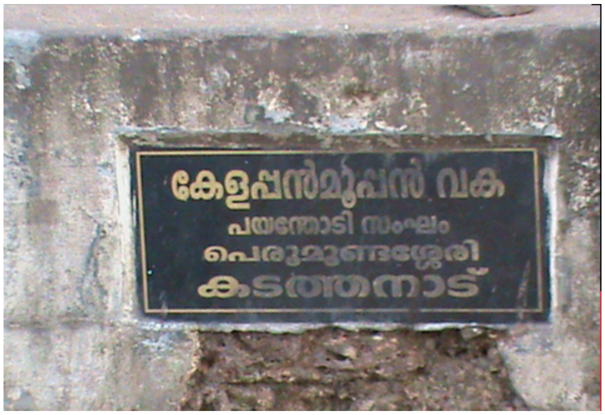
A photo-narrative was decided on as the best approach to convey the richness of the festival, chosen after abandoning other text-based approaches.

The photo-narrative had the advantage of being more accessible – both to the people I had sourced the content from, and to wider audiences.

Most Bharani participants speak Malayalam or Tamil. The crowd is mixed – literate and illiterate. While youngsters are familiar with English, the older people are not.

As I tried out various approaches to present my content, I realized that the photographs were the most evocative of all, and that weaving the narrative around those would help bridge literacy and language barriers, reaching out to young and old, local and global.

Excerpts



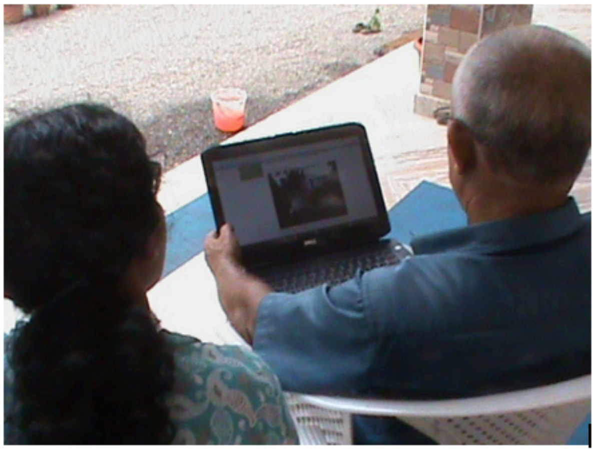
Signboard on the platform of one of the numerous trees that surround the temple, indicating the locality of the group to assemble at this site, (Picture: Own)

One of the groups assembled in 2014 (Picture: Own)

Validation

Once the presentation had reached a logical point, in 2016, I returned to the various persons I had obtained information from, for their validation and feedback. I also had independent knowledgeable individuals review the content, to ensure I had captured key terms in the narrative correctly.

The snapshots of the Bepur group viewing the site, are below. One of the youngsters of this group had asked me, when I was gathering my content – "What will be the result of your research? Will we get to see it, or will it sit in some library that we cannot access?"

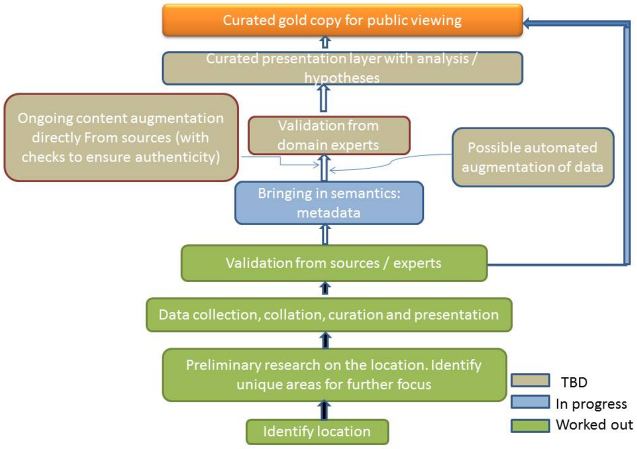


Although such an in-person validation was possible for the pilot site, as the system scales to multiple locations and contributors, as is the eventual goal, quantifiable and less resource-intensive methods of validation, using computational tools and systems as appropriate, will need to be implemented.

Future work

1. **Multi-Lingual Captions**
   The photo-narrative approach needs to be explored further, to include multi-lingual capability, especially in the narrative captions that will further increase accessibility.
2. **Semantics**
   Working out the conceptual categories and a suitable meta-data taxonomy to describe the content, will be crucial to make Limories a 'living' site. The diagram below depicts the overall process flow currently envisaged for the system.
3. **Scaling out**
   In parallel, strategies to scale out to other sites also need to be worked out. A second site has been chosen. The prospect of using local volunteers is also being explored.
4. **Social media as a source?**
   As Limories scales out, it 'might', in due course, use appropriate and relevant social media

content. However, a primary consideration, if so, would be appropriateness of content. A key intent of Limories is to reach out to all audiences, including children, for the younger generation to know the localities around them, and the associated memories they hold, better. Content would need to be accordingly appropriate.


Curated gold copy for public viewing
Curated presentation layer with analysis / hypotheses
Ongoing content augmentation directly From sources (with checks to ensure authenticity)
Validation from domain experts
Possible automated augmentation of data
Bringing in semantics: metadata
Validation from sources / experts
Data collection, collation, curation and presentation
Preliminary research on the location. Identify unique areas for further focus
Identify location
TBD
In progress
Worked out

Note: Chart to be read bottom-up

Working out the conceptual categories and meta-data, is the next major stage in the Limories journey, as the chart illustrates. This, and scaling out to other sites, with controlled crowd-sourcing, are crucial next steps for Limories, seminal to realizing its vision, to help answer: **What is it that makes a place unique and special, to all the people in whose memories the place resides?**

Bibliography

Varghese, M.A. (2013). "Spatial Reconfigurations and New Social Formations – The Contemporary Urban Context of Kerala" University of Bergen, Norway. ISBN:978-82-308-2285-2

Khair, T. (2016). **"**Questioning The Liberal Left" http://www.thehindu.com/opinion/op-ed/questioning-the-liberal-left/article9104473.ece (last accessed 7 Apr 2017)

Singh, N. (2003). **"**Nature of Historical Thinking and Aitihya.**"** *Studies in Humanities and Social Sciences*. X(2):1-28

Narayanan, M.G.S. (1972). "Perumals of Kerala." Ph.D. Dissertation published by Cosmobooks. ISBN 81-88765-07-4

Casson, L. (1989) "The Periplus Maris Erythraei. Text with Introduction, Translation and Commentary" Princeton University Press

Herbert, V. Classical Sangam poetry translation at http://sangamtranslationsbyvaidehi.com/ettuthokai-301-400/ **Puranānūru 343, Poet: Paranar** (last accessed: 7 Apr 2017)

Cherian P.J., Menon, Jaya.(2014) "Unearthing Pattanam: Histories, Cultures, Crossings" http://www.nationalmuseumindia.gov.in/pdfs/Pattanam-Catalogue-Masterlayout-05122014.pdf (last accessed: 7 Apr 2017)

Induchudan, V.T. (1969). "The Secret Chamber-A Historical, Anthropological and Philosophical Study of the Kodungallur Temple" The Cochin Devaswom Board

Gentes, M.J. (1992). "Scandalizing The Goddess At Kodungallur" *Asian Folklore Studies*, 1992(51):295-322.

Radhakrishnan, S (2014). "Sanitizing The Profane." http://subversions.tiss.edu/sanitizing-the-profane (last accessed: 7 Apr 2017)

Flexible NLP Pipelines for Digital Humanities Research

Janneke M. van der Zwaan
j.vanderzwaan@esciencecenter.nl
Netherlands eScience Center, The Netherlands

Wouter Smink
w.a.c.smink@utwente.nl
University of Twente, The Netherlands

Anneke Sools
a.m.sools@utwente.nl
University of Twente, The Netherlands

Gerben Westerhof
g.j.westerhof@utwente.nl
University of Twente, The Netherlands

Bernard Veldkamp
b.p.veldkamp@utwente.nl
University of Twente, The Netherlands

Sytske Wiegersma
s.wiegersma@utwente.nl
University of Twente, The Netherlands

Introduction

A lot of Digital Humanities (DH) research involves applying Natural Language Processing (NLP) tasks, such as, sentiment analysis, named entity recognition, or topic modeling. A large amount of NLP software is already available. On the one hand, there are frameworks that bundle software for different tasks and languages (e.g., NLTK [Bird et al, 2009], or xtas1), and on the other hand there are tools that target specific tasks (e.g., gensim, Rehurek and Sojka, 2010). As long as researchers do not need to combine tools from different packages, it is usually relatively easy to write scripts that perform the task. However, for innovative research, combining tools often is required, especially when working with non-English text. This abstract presents work in progress on NLP Pipeline (**nlppln**), an open source tool that improves access to NLP software by facilitating combining NLP functionality from different software packages2.

nlppln is based on Common Workflow Language (CWL), a standard for describing data analysis workflows and tools

(Amstutz et al, 2016). The main advantage of using a standard is that any existing NLP tool can be integrated into a workflow, as long as it can be run as a command line tool. This flexibility is missing from existing frameworks for creating NLP pipelines, such as DKPro (Eckart de Castilho, and Gurevych, 2015) using the UIMA framework (Ferrucci, and Lally, 2004). In addition to improved reuse of existing software, CWL increases research reproducibility, as it provides a standardized, formal record of all steps taken in a processing pipeline. Finally, CWL workflows are modular. This means that individual processing steps can easily be swapped in and out.

To demonstrate how NLP tools can be combined using nlppln, we show what need to be done to create a pipeline that removes named entities from a directory of text files. This is a common NLP task, that can be used as part of a data anonymization procedure.

The Software

An NLP pipeline or workflow is a sequence of natural language processing steps. A 'step' represents a specific NLP task, that is executed by a single tool. Tools require input and produce output. The basic units in CWL are command line tools (i.e., tools that can be run from the command line). In order to be able to run a command line tool, CWL needs a specification. The **nlppln** software helps creating those specifications. In addition, **nlppln** provides functionality to convert existing NLP tools written in Python to command line tools. Finally, the software helps users to combine (existing and new) processing steps to pipelines.

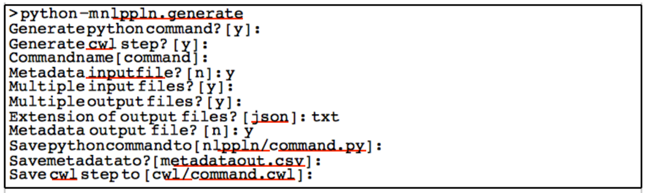
In the next section, we explain how **nlppln** facilitates creating NLP steps, and in "Constructing Pipelines" we demonstrate the creation of an NLP pipeline for data anonymization.

Generating Steps

nlppln allows users to generate CWL specifications for existing NLP tools. To simplify the generation of CWL specifications, we use a convention for NLP tasks. The convention assumes that there can be two types of input parameters: a list of files for which the command should be executed, and/or a file containing metadata about the texts in the corpus. Output parameters consist of a directory where output files are stored (usually there is one output file for every input file) and/or a file in which metadata is stored. So far, almost all steps that are currently available in **nlppln** follow this convention. Be that as it may, we would like to emphasize that it is possible to deviate from this convention; for example, when existing NLP functionality requires different parameters (e.g., the name of a directory containing the input files instead of a list of input files). This does however mean that the user has to adapt the CWL specification by hand.

In addition to CWL specifications, **nlppln** allows users to generate boilerplate Python command line tools. A boilerplate command line tool contains generic functionality, such as opening input files and saving output files, but lacks implementation of the specific NLP task. The generated Python command can be used to turn existing NLP functionality into command line tools, and to create Python command line tools for new NLP tasks.

Python commands and associated CWL steps are generated using a command line tool that requires the user to answer a sequence of yes/no questions. Listing 1 shows what that looks like for a (hypothetical) command 'command', that takes as input a metadata file and multiple input files, and produces as output multiple text files and metadata.



```
>python -m nlppln.generate
Generate python command? [y]:
Generate cwl step? [y]:
Commandname [command]:
Metadata inputfile? [n]: y
Multiple input files? [y]:
Multiple output files? [y]:
Extension of output files? [json]: txt
Metadata output file? [n]: y
Savepythoncommandto [nlppln/command.py]:
Savemetadatato? [metadataout.csv]:
Save cwl step to [cwl/command.cwl]:
```

Listing 1: Generating a CWL specification and associated boilerplate Python command using **nlppln.**

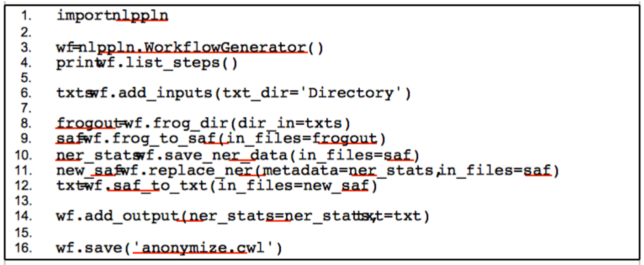
Constructing Pipelines

To combine text processing steps into a CWL pipeline, **nlppln** provides an interface that allows users to write a simple Python script. We demonstrate this functionality by creating a pipeline that replaces named entities in a collection of text documents. Named entities are objects in text referred to by proper names, such as persons, organizations, and locations. In the example pipeline, named entities will be replaced with their named entity type (i.e., PER (person), ORG (organization), LOC (location), or UNSP (unspecified)). The pipeline can be used as part of a data anonymization procedure.

The pipeline consists of the following steps:

1. Extract named entities from text documents using frog (van den Bosch et al, 2007), an existing parser/tagger for Dutch
2. Convert frog output to SAF, a generic representation for text data3
3. Aggregate data about named entities that occur in the text files
4. Replace named entities with their named entity type in the SAF documents
5. Convert SAF documents to text

All steps required for this pipeline are available through **nlppln**. Listing 2 shows the script that creates a CWL workflow for this pipeline. After importing **nlppln** (line 1), a new WorkflowGenerator object is created (line 3), and the available NLP steps are listed (line 4). Next, the script specifies the workflow inputs (line 6). In this case, there is a single input, which is a directory containing text files. This directory is the input of the first step, which is frog_dir (line 8). The output argument txts contains the internal CWL name of the input parameter (line 6). By assigning its value to the input argument dir_in of frog_dir (line 8), the output is connected to the input. Steps 1 to 5 from the pipeline description correspond to lines 8 to 12 in listing 2. After the remaining steps steps of the workflow are added (lines 9–12), the workflow outputs are specified (line 14). Finally, the workflow is saved to a CWL file (line 16).



```
1.   importnlppln
2.
3.   wf=nlppln.WorkflowGenerator()
4.   printwf.list_steps()
5.
6.   txts=wf.add_inputs(txt_dir='Directory')
7.
8.   frogout=wf.frog_dir(dir_in=txts)
9.   saf=wf.frog_to_saf(in_files=frogout)
10.  ner_stats=wf.save_ner_data(in_files=saf)
11.  new_saf=wf.replace_ner(metadata=ner_stats,in_files=saf)
12.  txt=wf.saf_to_txt(in_files=new_saf)
13.
14.  wf.add_output(ner_stats=ner_stats,txt=txt)
15.
16.  wf.save('anonymize.cwl')
```

Listing 2: Python script for constructing the pipeline to replace named entities in text files.

Conclusion

To help DH researchers to (re)use and combine existing NLP software, we presented **nlppln**, an open source Python package for creating flexible and reusable NLP pipelines in CWL. nlppln comes with ready-to-use NLP steps, facilitates creating new steps, and helps combining steps into standardized workflows that are portable across different software and hardware environments. Compared to existing frameworks for creating NLP pipelines, CWL and **nlppln** add flexibility and improved research reproducibility.

nlppln is a work in progress. An important challenge that needs to be addressed is the fact that there is no standard data format for representing text and/or information extracted from text. This means that we will have to add NLP steps that convert different data formats (cf. Eckart de Castilho, 2016)). For future work, we plan to implement additional NLP steps and pipelines, including functionality that targets more languages. We would also like to add visualizations of pipelines and allow users to run pipelines directly from **nlppln**.

Bibliography

Amstutz, P., Crusoe, M. R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., Scales, M., Soiland-Reyes, S., and Stojanovic, L. (2016). *Common Workflow Language, v1.0,*.

Bird, S., Loper, E., and Klein, E. (2009) *Natural Language Processing with Python*. O'Reilly Media Inc.

van den Bosch, A., B Busser, B., Dealemans, G. J., and Canisius, S. (2007) An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands*, pages 191–206, 2007.

Eckart de Castilho, R. (2016). Interoperability = f(community, division of labour). In *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 24–28, 2016.

Eckart de Castilho, R., and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014,* pages 1–11.

Ferrucci, D., and Lally., A. (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering* 10.3-4, pages 327–348.

Rehurek, R., and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Towards a Digital Narratology of Space

Gabriel Viehhauser-Mery
viehhauser@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Florian Barth
florianbarth@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Introduction

Besides time, characters and plot, space is one of the main components in storytelling. But despite its importance as a category for the setting of narrative action and unlike the other mentioned categories, the conceptualisation of space has long been neglected in narratological research. This holds true even after the so-called *spatial turn* (Soja 1990) in cultural history, that lead to a renewed interest and to fruitful insights into space as a metaphorical concept. However, a systematic description of the means by which space is created in narratives is still in its beginnings (e.g. Dennerlein 2009, Piatti et. al. 2009).

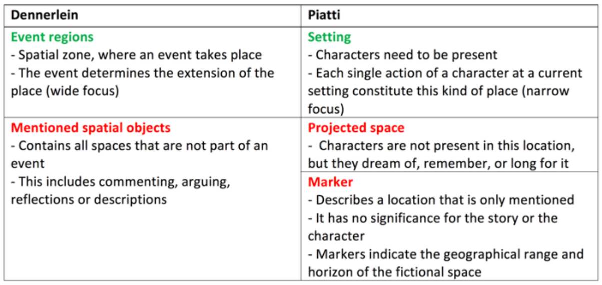
This is at least partly due to the fact that space poses substantial problems for modeling. The creation of space in narratives is often dynamic and based on implicit information: Rather than constructing a given, mathematical space beforehand, stories tend to evolve their setting in relation to its characters that constitute space through their actions. Spatial information in stories therefore highly depends on the characters that act, move or perceive within it. Especially in fiction, this also means that spatial information is often fuzzy and imprecise (Piatti et. al. 2009), since narrators quite frequently are more interested in telling a story than designing a detailed, coherent setting for it. Whereas these problems are hard to handle in traditional literary studies, they present serious yet interesting challenges for a digital formalization.

In our paper, we will illustrate the complex tasks that have to be tackled by a digital narratology of space based on an exemplary annotation workflow, that we will outline for the description of spatial elements in Jules Verne's *Around the world in Eighty Days.*

Challenges for a digital narratology of space

We describe the following major problems that can be grouped into a chain of work-tasks:

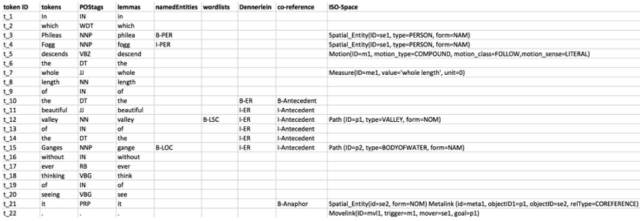
1. Basic information on the setting of a narrative can be retrieved by extracting the place names from a text (NER). However, an automatic extraction is flawed by well-known problems of disambiguation (cf. for example, Barbaresi / Biber 2016). Since space in narration is highly dependent on characters that are placed in it, these entities have to be detected as well.
2. Place names are not the only kind of spatial information that can be found in texts. Besides others, space is also constituted with the help of nouns that do not necessarily have an inherent spatial component (for example, a *car* in a text can be the subject of a description, but it turns into a space marker if someone enters it).
3. Spatial entities (names and nouns) can be referred to by co-reference.
4. Spatial entities in a text are not always the setting of narrative actions. Place names or nouns can also only be mentioned, dreamed of, remembered, reflected on etc. This different functionality has to be taken into account when it comes to the automatic generation of literary maps (eg. Moretti 1998, Piatti 2008). To capture this opposition, Dennerlein (2009) separates *event regions* from *mentioned spatial objects* in her conception. *Event regions* are defined as spatial zones, where events take place. In contrast, *mentioned spatial objects* contain all spaces that are not event-related. Piatti et. al. (2009) develop a similar model: Their concept of *setting* closely corresponds with *event regions*, whereas *projected space* and *marker* give a finer differentiation of the notion of *mentioned spatial objects* (cf. Figure 1).



Dennerlein	Piatti
Event regions - Spatial zone, where an event takes place - The event determines the extension of the place (wide focus)	Setting - Characters need to be present - Each single action of a character at a current setting constitute this kind of place (narrow focus)
Mentioned spatial objects - Contains all spaces that are not part of an event - This includes commenting, arguing, reflections or descriptions	Projected space - Characters are not present in this location, but they dream of, remember, or long for it
	Marker - Describes a location that is only mentioned - It has no significance for the story or the character - Markers indicate the geographical range and horizon of the fictional space

Figure 1: Comparison of spatial concepts from Katrin Dennerlein Barbara Piatti

Annotation Workflow



token ID	tokens	POStags	lemmas	namedEntities	wordlists	Dennerlein	co-reference	ISO-Space
t_1	In	IN	in					
t_2	which	WDT	which					
t_3	Phileas	NNP	philea	B-PER				Spatial_Entity(ID=se1, type=PERSON, form=NAM)
t_4	Fogg	NNP	fogg	I-PER				Spatial_Entity(ID=se1, type=PERSON, form=NAM)
t_5	descends	VBZ	descend					Motion(ID=m1, motion_type=COMPOUND, motion_class=FOLLOW,motion_sense=LITERAL)
t_6	the	DT	the					
t_7	whole	JJ	whole					Measure(ID=me1, value='whole length', unit=0)
t_8	length	NN	length					
t_9	of	IN	of					
t_10	the	DT	the			B-ER	B-Antecedent	
t_11	beautiful	JJ	beautiful			I-ER	I-Antecedent	
t_12	valley	NN	valley		B-LSC	I-ER	I-Antecedent	Path (ID=p1, type=VALLEY, form=NOM)
t_13	of	IN	of			I-ER	I-Antecedent	
t_14	the	DT	the			I-ER	I-Antecedent	
t_15	Ganges	NNP	gange	B-LOC		I-ER	I-Antecedent	Path (ID=p2, type=BODYOFWATER, form=NAM)
t_16	without	IN	without					
t_17	ever	RB	ever					
t_18	thinking	VBG	think					
t_19	of	IN	of					
t_20	seeing	VBG	see					
t_21	it	PRP	it				B-Anaphor	Spatial_Entity(id=se2, form=NOM) Metalink (id=meta1, objectID1=p1, objectID=se2, relType=COREFERENCE)
t_22	.	.	.					Movelink(ID=mvl1, trigger=m1, mover=se1, goal=p1)

The complexity of spatial information demands for a multi-faceted approach. Figure 2 shows a spreadsheet with the beginning of chapter 14 of Verne's *Around the World in Eighty Days* that has automatically, semi-automatically and manually been enriched with multiple layers of annotation.

First, the text was tokenized, lemmatized and POS-tagged (columns 1-4).

Second, Named Entity Recognition (NER) was applied to the text (columns 5, both steps were performed with Weblicht [2012]). The NER also identifies the names of the characters. The results of the NER have to be corrected manually. To improve the automation of this step the exploitation of other toponymical bases like Geonames and OpenStreetMap will be discussed.

Thirdly, to generate column 6, we used theme-specific wordlists that we built on the base of existing lexicological ontologies (*GermaNet* for German Texts [Hamp/Feldweg 1997, Henrich/Hinrichs 2010]. English Wordlists as shown were provisionally generated manually, but can be built in a similar way). These wordlists were used to automatically tag the text. In Fig. 1, 'valley' has been annotated as *LSC*, which means that it belongs to the word field *landscape*. So far, we created wordlists for *landscape* and *architecture* (in German), which cover a high amount of place nouns. The annotation can be manually supplemented and new wordlists can be created.

Fourthly, in column 7, 'the beautiful valley of the Ganges' was (manually) annotated as *event region* according to the model of Dennerlein (2009). An automatic differentiation between *event regions* and *mentioned spatial objects* will be a challenging task. However, we consider a rule-based extraction of dependency paths to approach the problem

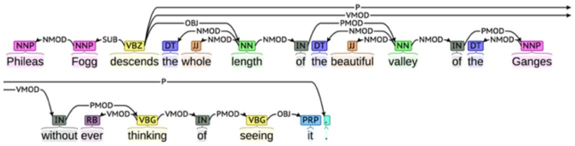


Figure 3 shows a parse tree of the sentence from Verne's text. The pattern [Character – SUBJ – Verb of Motion – OBJ – place noun] is likely to indicate an *event region*. By gaining several similar patterns with high precision regarding the identification of *event regions*, we assume to collect features for a future implementation of machine learning methods.

Fifthly, in column 8, coreferences were annotated manually. We will consider different kinds of co-references: Spatial entities can be referred to by nouns (e.g. 'Paris / 'city') or pronouns. Also, certain deictics ('here', 'there') might refer to spatial antecedents. However, a reliable automatic coreference resolution, which would be highly desirable for many kinds of narratological analysis, is out of the scope of this paper.

Finally, we included a column (9) for annotations that are based on the modified tag-set of the ISO-Space-standard (Pustejovsky et. al. 2011a und b), as they are presented in the SpaceEval Annotation Guidelines (2014).

Visualizations and Outlook

With the help of this semi-automatic and multi-layered method, we hope that we can make use of the strength of the different approaches and combine their advantages (it would be highly beneficial, for example, to combine Dennerlein's category with an ISO-space-annotated text to enhance their rule-based detection).

The potential of their combination shall be demonstrated by two examples of visualizations that draw on named entities and wordlists.

A network of spatial markers

As Piatti et. al. (2009) pointed out, the impreciseness and semantic potential of spatial information in literary texts sometimes demand visualizations other than geographical maps.

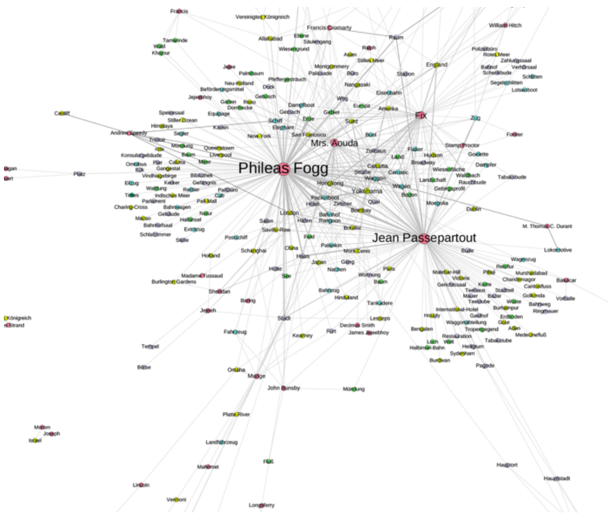


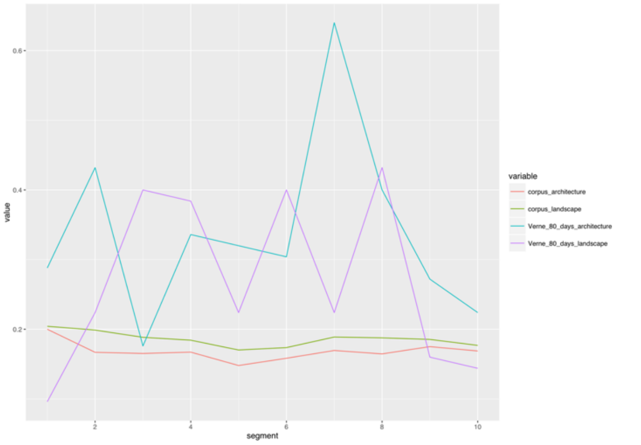
Figure 4 shows a co-occurrence network of characters and place markers in *Around the world in Eighty Days*: Characters appear in red, place names in yellow. Place nouns are divided into the sub-categories *landscape* (green), *architecture* (grey) and *transport* (blue). In a straightforward approach, we established edges whenever a character and a spatial marker appear in the same sentence. The nodes have been sized according to their degree (the number of their connections), which can be related to Juri Lotman's (Lotman 1977) concept of mobile vs. immobile characters: Characters which are connected to many places like Phileas Fogg and Passepartout are more likely to be main characters than characters with a lesser degree.

(The visualization was established with Gephi [Bastian et al. (2009)]).

Word-list-based Frequency Analyses

Figure 5 shows the distribution of *landscape* and *architecture* terms over the whole text of *Around the world in Eighty days* compared to a corpus of 451 German novels taken from the TextGrid-Repository (TextGrid Konsortium 2006-2014, licensed CC-BY-4.0), which cover a time range from 1700 to 1920.

Every text was chunked into 10 'segments' (x-axis), for which we calculated the relative percentage of the vocabulary from the corresponding word field ('value', y-axis). The graph shows a noticeable peak in the use of architectural vocabulary in the last third of the text, which can serve as a starting-point for a *close reading* of the text. However, to take full advantages of *distant reading* techniques for spatial analysis, more refined methods and annotated corpora are necessary. We hope that these methods can be developed by considering the challenges outlined in our basic model in this paper.



Bibliography

Barbaresi, A., and Biber, H. (2016): „Extraction and Visualization of Toponyms in Diachronic Text Corpora.“, in: Digital Humanities 2016, Jul 2016, Cracovie, Poland, Digital Humanities 2016 Conference Abstracts, 732-734 http://dh2016.adho.org/

Bastian, M., Heymann, S., and Jacomy, M. (2009): *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

Dennerlein, K.(2009): *Narratologie des Raumes*. Berlin: de Gruyter.

Hamp, B., and Feldweg, H.(1997): „GermaNet - a Lexical-Semantic Net for German.“, in: Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid

Henrich, V., and Erhard, H. (2010): "GernEdiT - The GermaNet Editing Tool", in: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta, May 2010, 2228-2235.

Lotman, J. (1977): *The Structure of the Artistic Text. Translated from the Russian by Ronald Vroon.* Ann Arbor: University of Michigan, Department of Slavic Languages and Literatures.

Moretti, F. (1998): *Atlas of the European novel. 1800-1900*. London / New York: Verso.

Piatti, B. (2008): *Die Geographie der Literatur. Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.

Piatti, B., Bär, H. R., Reuschel, A.-K., Hurni, L., Cartwright, W. (2009): „Mapping Literature: Towards a Geography of Fiction.", in: Cartwright, William / Gartner, Georg / Lehn, Antje (Eds.): Cartography and Art. Berlin / Heidelberg, Springer 2009, 179-194.

Pustejovsky, J., Moszkowicz, J. L., Verhagen, M. (2011a): ISO-Space: The annotation of spatial information in language, in: Proceedings of the Joint ACL-ISO Workshop on Interoperable Semantic Annotation, 1–9

Pustejovsky, J., Moszkowicz, J. L., Verhagen, M. (2011b): Using ISO-Space for Annotating Spatial Information. In: Proceedings of the International Conference on Spatial Information Theory

Soja, E. (1990): *Postmodern Geographies. The Reassertion of Space in Critical Social Theory.* London / New York: Verso.

SpaceEval Annotation Guidelines (2014) http://jamespusto.com/wp-content/uploads/2014/07/SpaceEval-guidelines.pdf

TextGrid Konsortium (2006–2014). *TextGrid: Virtuelle Forschungsumgebung für die Geisteswissenschaften.* Göttingen: TextGrid Konsortium. textgrid.de.

WebLicht (2012): CLARIN-D/SfS-Uni. Tübingen 2012. WebLicht: Web-Based Linguistic Chaining Tool. Online. https://weblicht.sfs.uni-tuebingen.de/

Evaluating Research Practices in the Digital Humanities by Means of User Activity Analysis

Niels-Oliver Walkowski
walkowski@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften Germany

Approaches to the Evaluation of DH Research Processes

The documentation of digital research processes has been a heavily discussed topic for some years now. It is most often addressed by the term *provenance*. In most cases, provenance data is created to record the chain of production of digital research results, in order to increase transparency in research and make such results reproducible.

In more recent times, a slightly different version of this topic has appeared in the field of digital humanities. Accordingly, digital research processes are modeled and documented with the aim to identify methodologies and practices of digital research in the arts and humanities on a broader scale. Two models have been introduced in this respect. One is the *Scholarly Domain Model* (SDM) (Gradmann et al, 2015) the other is the *NeDiMAH Method Ontology* (NeMO) (Constantopoulos, Dallas, and Bernadou, 2016). Both models provide formal semantics for the description of research processes as well as for their methodological evaluation.

The project environment in which these models were defined was dominated by European infrastructure projects, specifically *DARIAH* in the case of NeMO and *Europeana* in the case of SDM. Accordingly, such models aim to identify user needs and the qualitative use of infrastructure services as a first priority. However, it is also easily possible to refer to them in a broader perspective of *laboratory research* and *science studies*. For the case of the digital humanities community such a perspective corresponds with this community's wish to develop a unique methodological self-awareness.

Upon closer observation, the two models take up a different approach to accomplish their goals. In terms of NeMo, applications of the model describe research processes after they have taken place. In contrast, SDM makes reference to the concept of "modeling for" by Clifford Geertz and describes research processes in advance.

Such difference calls for a more concise evaluation of the terminology that was used before. In the research literature three concepts are used to distinguish between three possible viewpoints from which research processes can be described (Hunter 2006). These concepts are: workflow, provenance and lineage. As it has been indicated in the evaluation of SDM and NeMO the difference of these viewpoints is marked by the place in time from which they look at a research process. As such, it is also possible to call these concepts "prescribing", "inscribing", and "describing".

In accordance with this systematization SDM defines workflows while NeMO presents lineages. What is missing however, is real provenance data that is created and modeled in order to systematically evaluate digital humanities practices. More specifically, this means data which is recorded during the research process. The main goal of this presentation is to introduce an approach for how such provenance data can be created and modeled meaningfully to reach its goal by way of example. The example is the Wissensspeicher at the Berlin-Brandenburg Academy of Sciences and Humanities. The Wissensspeicher connects all digital resources of the academy in a way that lets the user interact not just with metadata but with parts of the content itself. The work which will be presented is part of the evaluation phase of the DARIAH-DE project that started in March 2016.

When evaluating research practices provenance data has some advantages in comparison to workflow or lineage data. On the one hand, it is easier to obtain very detailed data. On the other hand, semantic implications which might predefine the results of the evaluation are less necessary. For instance, in the SDM primer the concept of annotating exists before it is applied to an activity in a specific situation. In consequence, research results about digital annotation practices are biased. They depend on personal decisions of someone who classifies activities as annotating, or on a normative concept of annotating. For the identification of new research practices in digital environments, both aspects are problematic. Provenance data does not have the same risk because it is mostly created before classification takes place. The only aspect which is predefined is the fact that recordable actions take place and that these actions form part of a broader research context.

Nevertheless, the implementation of technological solutions for the creation of provenance data in these circumstance is more complicated than in common situations where provenance data is created. It is not enough to record which software component modifies a digital resource at a certain point of time as it happens in e-science. The "resource" is the research process itself, the actions take place on multiple levels and such actions are carried out not only by software but also by humans.

User Activity Analysis and Digital Humanities Research Processes

In fact, there is one field of research which addresses a comparable situation and this field is **user activity analysis.** In user activity analysis, human-computer-interaction is recorded in order to evaluate the user with respect to a specific research interest. The approach is used in areas like **e-commerce** and **online social networks** research in order to create services like recommendation systems (Plumbaum, Stelter, and Korth 2009) or to analyze social behavior (Dang et al. 2016). There are few examples of user activity analysis in academic digital environments. Suire et al. (2016) use this approach in the cultural heritage domain while Vozniuk et al. (2016) applies it to model learning processes in e-learning environments.

Having said that, no ready-made solution exists which can be easily used in the present context. Instead different approaches to user-activity analysis have to be evaluated in order to decide which ones can be adopted. Nevertheless, under the circumstances of evaluating digital research practices these decisions remain contingent. Digital research takes place in very different digital environments and under different conditions. Thus, in every situation in which digital research practice should be evaluated a different selection from the existing set of options might be the best. An overview of these options will be published in a DARIAH-DE report in the future.

The advantage of the Wissensspeicher use-case is the fact that it is a web platform– most user activity analysis takes place on websites and in web environments. There are two major tasks which need to be distinguished. The first task is user activity tracking and concerns how the data is created. The second task is the actual analysis. It demands to evaluate in which sense the created data constitute meaningful events and how to make sense out of these events.

Use–Case: Wissensspeicher

The Wissensspeicher implements user activity tracking by combining three different strategies: http-request logging, browser-event parsing and user annotations. Http-requests are logged by virtue of the *Django* request object and the *logger* library in the Python Django app that creates the website. Thereby request information can be pre-processed when it is detected. When a page is loaded in the browser a *JavaScript* client registers event listeners for page elements and certain user actions. Each event that is triggered causes the client to parse relevant information in the *DOM* of the *HTML* including microdata which has been created in the Django app in advance. Additionally, the user is able to directly give feedback in some situations. The created data is stored in a *MongoDB* database.

User activity analysis is also realized by virtue of three steps. In a certain way these steps resemble the three angles of workflow, provenance and lineage. First, events are evaluated in a so called *task model*. This task model describes ideal sequences of actions and user goals as conceived by the project employees. Second, users are evaluated by applying the **thinking aloud** technique from the field of **usability testing.** Finally, existing data will be evaluated to identify common event sequences by computing its clusters. A systematization of the results from these evaluations will enables researchers to associate certain meanings with events in such a way that the data can be analyzed to permit insights into research practices within the use case.

A Dialogue of Approaches

This presentation will summarize activities to evaluate research practices and methods in the digital humanities. It will outline a unique and complementary approach and indicate how this approach can be used in conjunction with existing digital humanities research practices. Finally, the implementation and results will be described up to the point that such results are present after two-thirds of the project time has elapsed.

Bibliography

Constantopoulos, P., Dallas, C., and Bernadou, A. (2016). "Digital Methods in the Humanities: Understanding and Describing Their Use Across the Disciplines." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 1st ed. Chichester, West Sussex, UK: John Wiley & Sons.

Dang, A., Moh'd, A., Milios, E., and Minghim, R. 2016. "What Is in a Rumour: Combined Visual Analysis of Rumour Flow and User Activity." In *Proceedings of the 33rd Computer Graphics International*, 17–20. ACM.

Gradmann, S., Hennicke, S., Tschumpel, G., Dill, K., Thoden, K., Pichler, A., and Morbidoni, C. (2015). "Beyond Infrastructure! Modelling the Scholarly Domain."

Hunter, J. (2006). "Scientific Models: A User-Oriented Approach to the Integration of Scientific Data and Digital Libraries." In *VALA2006*, 1–16.

Plumbaum, T., Stelter, T., and Korth, A. (2009). "Semantic Web Usage Mining: Using Semantics to Understand User Intentions." In *User Modeling, Adaptation, and Personalization*, edited by Geert-Jan Houben, Gord McCalla, Fabio Pianesi, and

Massimo Zancanaro, 391–96. Lecture Notes in Computer Science 5535. Springer Berlin Heidelberg. doi:*10.1007/978-3-642-02247-0_42*.

Suire, C., Jean-Caurant, A., Courboulay, V., Burie, J.-C., and Estraillier, P. (2016). "User Activity Characterization in a Cultural Heritage Digital Library System." In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 257–58. ACM.

Vozniuk, A., Rodriguez-Triana, M.J., Holzer, A, and Gillet, D. (2016). "Combining Content Analytics and Activity Tracking to Identify User Interests and Enable Knowledge Discovery." In *Proceedings of the 6th Workshop on Personalization Approaches in Learning Environments (PALE 2016)*. 24th conference on User Modeling, Adaptation, and Personalization (UMAP 2016), CEUR workshop proceedings, this volume.

Collaborative Writing to Build Digital Humanities Praxis

Brandon Walsh
bmw9t@virginia.edu
University of Virginia, United States of America

Introduction

Emergent programs like those associated with the Praxis Network have redefined the possibilities for digital humanities training by offering models for project-based pedagogy. These efforts provide innovative institutional frameworks for building up and sharing digital skills, but they primarily focus on graduate education. The long-term commitments that they require can make them difficult to adapt for the professional development of other librarians, staff, and faculty collaborators. While members of these groups might share deep interests in undertaking such programs themselves, their institutional commitments often prevent them from committing the time to such professional development, particularly if the outcomes are not immediately legible for their own structures of reporting.

My talk argues that we can make such praxis programs viable for broader communities by expanding the range of their potential outcomes. This talk explores the potential for collaborative writing projects to develop individual skillsets and, by extension, the capacity of digital humanities programs. While the example here focuses on a coursebook written for an undergraduate audience, I believe the model and set of pedagogical issues can be extrapolated to other circumstances. By considering writing projects as potential opportunities for project-based development, I argue that we can produce professionally legible outcomes that both serve institutional priorities and prove useful beyond local contexts.

Case Study

The particular case study for this talk will be an open coursebook written for a course on digital text analysis (Walsh and Horowitz, 2016). In the fall of 2015, Professor Sarah Horowitz, a colleague in the history department at Washington and Lee University, approached the University Library with an interest in digital text analysis and a desire to incorporate these methods in her upcoming class. As the Mellon Digital Humanities Fellow working in the University Library, I was asked to support Professor Horowitz's requests because of my own background working with and teaching text analysis. Professor Horowitz and I conceived of writing the coursebook as a means by which the Library could meet her needs while also building the capacity of the University's digital humanities resources. Our model in this regard was as an initiative undertaken by the Digital Fellows at the CUNY Graduate Center, where their Graduate Fellows produce documentation and shared digital resources for the wider community. We aimed to expand upon their example, however, by making collaborative writing a centerpiece of our pedagogical experiment. Through her involvement in the the creation of the course materials, Professor Horowitz engaged with a variety of technologies: Markdown, Git, and GitHub. The process also required synthesis of both text analysis techniques and disciplinary material relevant to a course in nineteenth-century history. As a result of our initial collaboration in writing the materials and teaching the course, Professor Horowitz is prepared to offer the course herself in the future without the support of the library. In addition, we now possess course materials that could, after careful structuring and selection of platforms, be reusable in other courses at our own institution and beyond.

This type of writing collaboration can fit the professional needs of people in a variety of spaces in the university. Course preparation, for example, often takes place behind the scenes and away from the eyes of students and other scholars. With a little effort, the hidden labor of teaching can be transformed into openly available resources capable of being remixed into other contexts. As Shawn Graham (2016) has illustrated through his own resources for a class on Crafting Digital History, course materials can be effectively leveraged to serve a wider good in ways that still parse in a professional context. In our case, the collaboration produced public-facing web writing in the form of an open educational resource. The history department regarded the project as a success for its potential to bring new courses, skills, and students into the major as a result of Professor Horowitz's training. The University Library valued the collaboration for its production of open access materials, development of faculty skills, and exploration of workflows and platforms for faculty collaboration. We documented and managed the writing process in a GitHub repository. This versioned workflow was key to our conception of the project, as we hoped to structure the project in such a way that others could copy down and spin up their own versions of the course materials for their own needs. We were careful to compartmentalize the lessons according to their focus on theory, application, or course exercises, and we provided documentation to walk readers through the technical process of adapting the book to reflect their own disciplinary content.

Implications for DH Praxis

My talk argues that writing projects like this one provide spaces for shared learning experiences that position student and teacher as equals. By writing in public and asking students and faculty collaborators to discuss, produce, and revise open educational resources, we can break down distinctions between writer and audience, teacher and student, programmer and non-programmer. In this spirit, work by Robin DeRosa (2016) with the Open Anthology of Earlier American Literature and Cathy Davidson with HASTAC has shown that students can make productive contributions to digital humanities research at the same time that they learn themselves. These contributions offer a more intimate form of pedagogy – a more caring and inviting form of building that can draw newcomers into the field by way of non-hierarchical peer mentoring. It is no secret that academia contains "severe power imbalances" that adversely affect teaching and the lives of instructors, students, and peers (McGill, 2016). I see collaborative writing as helping to create shared spaces of exploration that work against such structures of power. They can help to generate what Bethany Nowviskie (2016) has recently advocated as a turn away from the Kantian ideal of an isolated, reasoning self and towards, instead, a "feminist ethics of care" to "illuminate the relationships of small components, one to another, within great systems." By writing together, teams engage in what Nowviskie (2011) calls the "perpetual peer review" of collaborative work. Through conversations about ethical collaboration and shared credit early in the process, we can privilege the voice of the learner as a valued contributor to a wider community of practitioners even before they might know the technical details of the tools or skills under discussion.

Collaborative writing projects can thus serve as training in digital humanities praxis: they can help introduce the skills, tools, and theories associated with the field, and projects like ours do so in public. Productive failure in this space has long been a hallmark of work in the digital humanities, so much so that "Failure" was listed as a keyword in the new anthology *Digital Pedagogy in the Humanities* (Croxall and Wernick, 2016). Writing in public carries many of the same rewards – and risks. While a certain comfort with frustration can help one learn digital methods (Ramsay, 2016) not everyone is comfortable with what Stephen Ramsay (2014) describes as a "hermeneutics of screwing around." Many of those new to digital work, in particular, rightfully fear putting their work online before it is published. I argue that the clearest way in which we can invite people into the rewards of public digital work is by sharing the burdens and risks of such work. In her recent work on generous thinking, Kathleen Fitzpatrick (2016) advocates "rooting the humanities in generosity, and in particular in the practices of thinking *with* rather than reflexively *against* both the people and the materials with which we work". By framing digital humanities praxis first and foremost as an activity whose successes and failures are shared, we can lower the stakes for newcomers. Centering this approach to digital humanities pedagogy in the practice of writing productively displaces the very digital tools and methodologies that it is meant to teach. Even if the ultimate goal is to develop a firm grounding in a particular digital topic, focusing on the writing invites students and collaborators into a space where anyone can contribute. By privileging soft rather than technical skills as the means of engagement and ultimate outcome, we can shape a more inviting and generous introduction to digital humanities praxis.

Bibliography

Croxall, B. and Warnick, Q. (2016). "Failure." In *Digital Pedagogy in the Humanities: Concepts, Models, and Experiments*. Modern Languages Association.

DeRosa, R. (2016). "The Open Anthology of Earlier American Literature." https://openamlit.pressbooks.com/

Fitzpatrick, K. (2016). "Generous Thinking: The University and the Public Good." *Planned Obsolescence*. http://www.plannedobsolescence.net/generous-thinking-the-university-and-the-public-good/

Graham, S. (2016). "Crafting Digital History." http://workbook.craftingdigitalhistory.ca/

McGill, B. (2016). "Serial Bullies: An Academic Failing and the Need for Crowd-Sourced Truthtelling." *Dynamic Ecology*. https://dynamicecology.wordpress.com/2016/10/18/serial-bullies-an-academic-failing-and-the-need-for-crowd-sourced-truthtelling/.

Nowviskie, B. (2016). "Capacity Through Care." http://nowviskie.org/2016/capacity-through-care/

Nowviskie, B. (2014). "The Hermeneutics of Screwing Around; or What You Do with a Million Books." In *Pastplay: Teaching and Learning History with Technology*, edited by Kevin Kee. University of Michigan Press. http://hdl.handle.net/2027/spo.12544152.0001.001

Nowviskie, B. (2011). "Where Credit Is Due. http://nowviskie.org/2011/where-credit-is-due/

Ramsay, S. (2010). "Learning to Program." http://stephenramsay.us/2012/06/10/learning-to-program/

The Praxis Network (2017). University of Virginia Library's Scholars' Lab. http://praxis-network.org/

Walsh, B. and Horowitz, S. (2016). "Introduction to Text Analysis: A Coursebook." http://www.walshbr.com/textanalysiscoursebook

Designing Collaborative Ecosystems and community organization: Introducing the multidisciplinary portal on "Biodiversity and Linguistic Diversity: A Collaborative Knowledge Discovery Environment"

Eveline Wandl-Vogt
eveline.wandl-vogt@oeaw.ac.at
Austrian Center for Digital Humanities
Austrian Academy of Sciences, Austria

Davor Ostojic
davor.ostojic@oeaw.ac.at
Austrian Center for Digital Humanities
Austrian Academy of Sciences, Austria

Barbara Piringer
barbara.piringer@oeaw.ac.at
Austrian Center for Digital Humanities
Austrian Academy of Sciences, Austria

Heimo Rainer
heimo.rainer@nhm-wien.ac.at
Natural History Museum Vienna, Austria

Ksenia Zsaytseva
ksenia.zsaytseva@oeaw.ac.at
Austrian Center for Digital Humanities
Austrian Academy of Sciences, Austria

The UNESCO Declaration on Cultural Diversity (United Nations, 2002) states the importance of Cultural Diversity as one of the "common heritage of humanity" and makes its defence "an ethical imperative indissociable from respect for the dignity of the individual." In declaring the hope, "that one day the declaration will acquire the same force as the Universal Declaration of Human rights", Koïchiro Matsuura, Director General, manifests the importance of the Declaration for global peace and one of the founding texts of the new ethics of the twenty-first century UNESCO (Wandl-Vogt et al, 2016).

There is a fundamental linkage between biological, linguistic and cultural diversity. "Over the past decade, the field of biocultural diversity has arisen as an area of transdisciplinary research concerned with investigating the links between the world´s linguistic, cultural, and biological diversity as manifestations of the diversity of life" (Maffi, 2005).

The interest goes back to the observation that all dimensions are under threat by some of the same forces. It was realized, that this may lead to dramatic consequences for humanity and the earth. "While it is widely acknowledged, that the degradation of the natural environment, in particular habitats, entails a loss of cultural and linguistic diversity, new studies suggest that language loss, in its turn, has a negative impact on biodiversity conversation. There is a fundamental linkage between language and traditional knowledge related to biodiversity. Local and indigenous communities have elaborated complex classification systems for the natural world, reflecting a deep understanding of their local environment. This environmental knowledge is embedded in indigenous names, oral traditions and taxonomies, and can be lost when a community shifts to another language" (UNESCO, n.d.). Several studies, mainly carried out on indigenous languages, demonstrate mutual benefits. UNESCO (see above) considers the safeguarding of traditional knowledge and indigenous as well as local languages used to transmit such knowledge as a yet "underused but promising tool for the conversation and sustainable management of biodiversity".

This paper introduces into a multidisciplinary, collaborative knowledge discovery environment (AAS and ACDH, 2016), established at the Austrian Academy of Sciences (AAS), Austrian Centre for Digital Humanities (ACDH) in close collaboration with Natural History Museum Vienna (NHM), launched on September, 1st, 2016.

The project aims to establish a collaborative open science workflow and enable knowledge discovery as well as experimental scholarship related to biodiversity, linguistic diversity and it´s cultural dimensions. It aims to serve as a social and technical infrastructure enabling research for a multidisciplinary community of practice. The portal contributes to the sustainable documentation and visibility of cultural diversity and traditional knowledge and to open it for the public.

In this paper we introduce into first results of our endeavor, yet focus on the design process and community building efforts concerning the "Biodiversity and Linguistic diversity portal" (diversity4bio).

diversity4bio started as a case-study-collaboration between Natural History Museum Vienna (NHM), Heimo Rainer, and Austrian Academy of Sciences (AAS), Eveline Wandl-Vogt, in 2012, in the framework of the European funded project "OpenUp! Opening up European Natural History Heritage for Europeana".

Based on the background of the partners (Taxonomy, Nonstandard-language Lexicography) and the OpenUp!-projects main focus (common names for Europeana metadata) the first step to get towards results was to work on linguistic diversity.

The partners designed the project as a community organization. Community organization in this paper is understood as organizing within communities defined by shared experience, shared curiosity and interests as well as shared (virtual) work space. Community organization is a process by which a community identifies needs and takes action, and in doing so develops co-operative attitudes and practices (Ross, 1955)

The main characteristics of Community organization are discussed on the given example (Wallin, 2016):

1. **The boundaries are fluid, informal:** Everybody who contributes to diversity4bio belongs to the network. There is no need of a certain scientific status or collaboration contract. Members support each other to find the right place to fit in. Still, it is visibility, citation and reliability is relevant for the network, which means a high degree on transparence for data interaction and publication.
2. **Significant incorporation of voluntary labor:** Community organization does not end in legal contracts or financial remuneration. diversity4bio recently is a loose network, where "members" are connected by common spirit and visions for further developments rather than legal contracts.
3. **Significant open sharing of knowledge:** diversity4bio is aiming to bring together experts to exchange and explore. The authors implement open science workflows based on the actors needs and the visons of the open science definition. It is introducing open science commons especially in research areas where these are still atypical, e.g. in traditional lexicography units. The digital transformation is described as a time taking process of building trust, where first sample data sets – e.g. names for living organisms – are implemented into new collaborative, open approaches, enriched and methodologically innovatively exploited.
4. **Collaborative communities display:** Behind the virtual portal scenery, there is a shared ethic of independent contribution and there is a formalized set of norms of interdependent process management as well as an interactive social character and identity. Community building was a process of several years, which is reflected in the presentation, e.g. face to face meetings of the protagonists, just after first results of collaboration (biological data + linguistic data > Europeana-metadata workflow) workshop to connect others and opening up of the group towards a portal and first publication of results.

The very interactive and flexible community organization model is certainly a challenge, especially as diversity4bio is mixing up several disciplines without a very certain or concrete research focus.

Based on this fact we do have established network facilitators for botanics/taxonomy, Heimo Rainer, and cultural data/lexicography, Eveline Wandl-Vogt.

Technically seen, diversity4bio is a first example for an open science workflow design. The architecture makes use of existing (freely available) technologies and connects with existing infrastructures.
The architecture consists of three layers:

1. The Human Interface Layer
2. The Persistent Layer
3. The Enrichment Layer

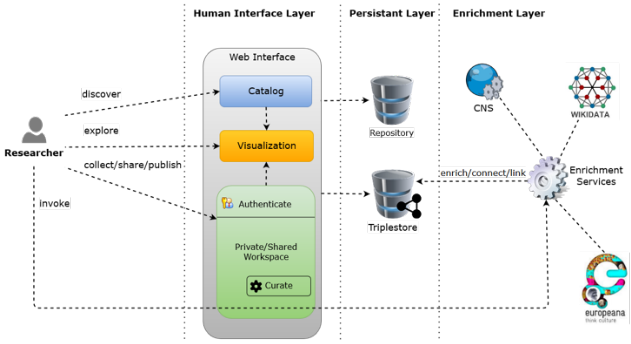
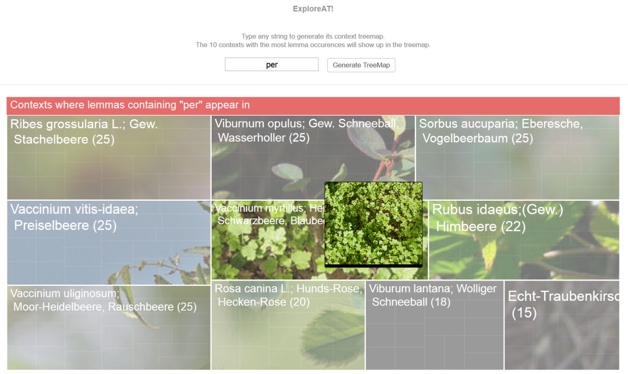


Figure 1: Pilot-architecture of the open workflow for the diversity4bio © Davor Ostojic

Data Visualization is part of the Human Interface Layer and developed within the project exploreAT! (exploring austria´s culture through the language glass). Furthermore, within the COST action IS1305 ENeL (European Network of electronic Lexicography) the interlinking and the use of the framework was stimulated. This may be seen as example on how ongoing initiatives contribute to the collaborative approach and vice versa.


ExploreAT!

Type any string to generate its context treemap.
The 10 contexts with the most lemma occurences will show up in the treemap.

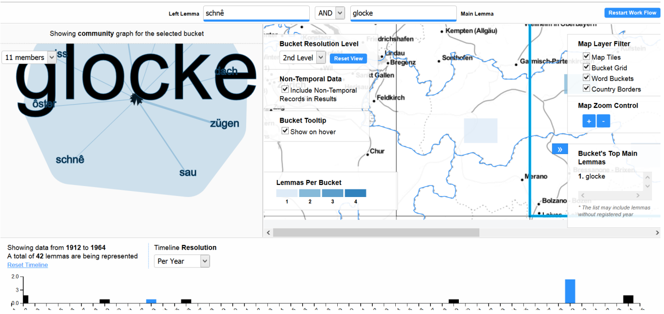
per

Generate TreeMap

Contexts where lemmas containing "per" appear in

Ribes grossularia L.; Gew. Stachelbeere (25)
Viburnum opulus; Gew. Schneeball, Wasserholler (25)
Sorbus aucuparia; Eberesche, Vogelbeerbaum (25)
Vaccinium vitis-idaea; Preiselbeere (25)
Vaccinium myrtillus; Hei Schwarzbeere, Blaube
Rubus idaeus;(Gew.) Himbeere (22)
Vaccinium uliginosum; Moor-Heidelbeere, Rauschbeere (25)
Rosa canina L.; Hunds-Rose, Hecken-Rose (20)
Viburum lantana; Wolliger Schneeball (18)
Echt-Traubenkirsc (15)

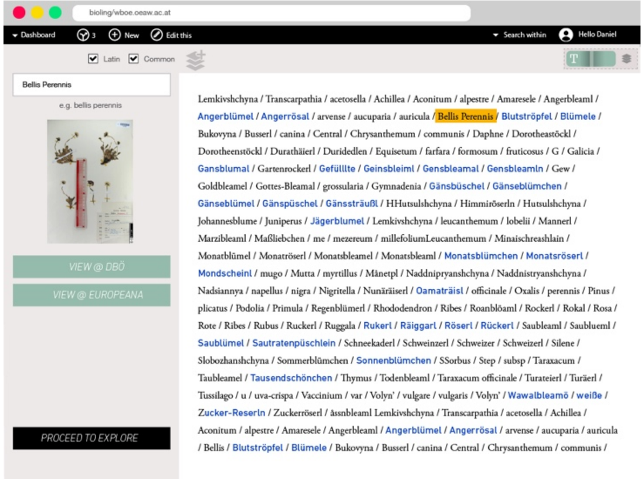
Figure 2,3 : Visualisation pilots developed in the project exploreAT!: Generalised and ready to be scaled up for global resources © Roberto Theron et al.



The portal offers a reconcile service. It queries the so called Common Names Service (CNS) at NHM for common names of living organisms in different languages for a given scientific name (Latin, Greek). The portal offers the opportunity of a CSV-import with the own data collection. Data import is time consuming. CNS itself is a distributed service.

Finally, the portal offers – as a first step – for a CSV-import of scientific names script-based RDF-modeling. Further enrichment services are in development, but at the time being not implemented into the portal, e.g. script-based semi-automatically interconnection with the Europeana – workflow mentioned above. Furthermore, out of the box repositories for researchers and citizens with less technical knowledge are necessary to keep in track with our vision to document traditional knowledge and related local language in development.

Finally, design probes for the collaborative process have been developed. The interlinked scientifically interpreted data may be reused to build a Pan-European (COST ENeL vision) or even global dictionary of plant names and envision based on this new movements for collaboration as well as research.



Lemkivshchyna / Transcarpathia / acetosella / Achillea / Aconitum / alpestre / Amaresele / Angerbleaml / Angerblümel / Angerrösal / arvense / aucuparia / auricula / Bellis Perennis / Blutströpfel / Blümele / Bukovyna / Busserl / canina / Central / Chrysanthemum / communis / Daphne / Dorotheastöckl / Dorotheenstöckl / Durathäierl / Durideedlen / Equisetum / farfara / formosum / fruticosus / G / Galicia / Gansblumal / Gartenrockerl / Gefülllte / Geinsbleiml / Gensbleamal / Gensbleamln / Gew / Goldbleamel / Gottes-Bleamal / grossularia / Gymnadenia / Gänsbüschel / Gänseblümchen / Gänseblümel / Gänspüschel / Gänssträußl / HHutsulshchyna / Himmiröserln / Hutsulshchyna / Johannesblume / Juniperus / Jägerblumel / Lemkivshchyna / leucanthemum / lobelii / Mannerl / Marzibleaml / Maßliebchen / me / mezereum / millefoliumLeucanthemum / Minaischreashlain / Monatblümel / Monatröserl / Monatsbleamel / Monatsbleaml / Monatsblümchen / Monatsröserl / Mondscheinl / mugo / Mutta / myrtillus / Mânetpl / Naddnipryanshchyna / Naddnistryanshchyna / Nadsiannya / napellus / nigra / Nigritella / Nunäräiserl / Oamaträisl / officinale / Oxalis / perennis / Pinus / plicatus / Podolia / Primula / Regenblümerl / Rhododendron / Ribes / Roanblöaml / Rockerl / Rokal / Rosa / Rote / Ribes / Rubus / Ruckerl / Ruggala / Rukerl / Räiggarl / Röserl / Rückerl / Saubleaml / Saublueml / Saublümel / Sautratenpüschlein / Schneekaderl / Schweinzerl / Schweizer / Schweizerl / Silene / Slobozhanshchyna / Sommerblümchen / Sonnenblümchen / SSorbus / Step / subsp / Taraxacum / Taubleamel / Tausendschönchen / Thymus / Todenbleaml / Taraxacum officinale / Turateierl / Turäerl / Tussilago / u / uva-crispa / Vaccinium / var / Volyn' / vulgare / vulgaris / Volyn' / Wawalbleamö / weiße / Zucker-Reserln / Zuckerröserl / åssnbleaml Lemkivshchyna / Transcarpathia / acetosella / Achillea / Aconitum / alpestre / Amaresele / Angerbleaml / Angerblümel / Angerrösal / arvense / aucuparia / auricula / Bellis / Blutströpfel / Blümele / Bukovyna / Busserl / canina / Central / Chrysanthemum / communis /

Figure 4: Design Probes for the further development of the diversity4bio portal towards a broader audience with cultural interest © Alisa Goikhman.

In conclusion, the authors give an outlook on next steps, pathways to contribute and join in and share the personal experiences along the way.

Acknowledgements

diversity4bio is connected to and partially financially supported by ongoing initiatives and projects, e.g. COST action IS 1305 European Network for electronic Lexicography (ENeL), exploreAT! (exploring austria´s culture through the language glass), DARIAH-EU, Europeana, EGI ENGAGE, SCI GAIA.

Bibliography

Austrian Centre for Digital Humanities (2016): Biodiversity and Linguistic Diversity: Collaborative Knowledge Discovery Environment: https://reconcile.eos.arz.oeaw.ac.at/ Austrian Academy of Sciences (AAS), (published: September 1st 2016; accessed: November, 1st 2016).

European Commission (2013): Open Innovation 2.0.

FOSTER Open Science. (n.d.) Open Science Definition: https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition (accessed: November, 1st 2016).

Maffi, L. (2005): Linguistic, Cultural, and Biological Diversity. Annual Review of Anthropology. Vol. 34: 599-617. DOI: 10.1146/annurev.anthro.34.081804.120437

OpenUp! (n.d.) OpenUp! Opening up the Natural History Heritage for Europeana. http://open-up.eu/en (accessed: November, 1st 2016).

Ross, M. G. (1955). Community Organization. New York. Community organization is a process by which a community identifies needs and takes action, and in doing so... develops co-operative attitudes and practices (Murray G. Ross, 1967) nach Wikipedia (https://en.wikipedia.org/wiki/Community_organization). (accessed: November, 1st 2016).

UNESCO. (n.d.) Biodiversity and linguistic diversity. Maintaining indigenous languages, conserving biodiversity. http://www.unesco.org/new/en/culture/themes/endangered-languages/biodiversity-and-linguistic-diversity/ (accessed: November, 1st 2016).

United Nations (2002): Universal Declaration of Cultural Diversity. A vision, a conceptual platform, a pool of ideas for implementation, a new paradigm. Cultural Diversity Series 1. http://unesdoc.unesco.org/images/0012/001271/127162e.pdf (accessed: November, 1st 2016)

Wallin, M. (2016; unpublished): Lab for Open Innovation in Science (LOIS). Module 11: Ecosystems for Open Innovation in Science. PPT. October, 21st-22nd 2016.

Wandl-Vogt, E., Barbera, R., Davidovic, D., et al (2016): Furthering the Exploration of Language Diversity and Pan-European Culture: The DARIAH-CC Science Gateway for Lexicographers. In: Margalitadze, Tinatin et al (Eds., 2016): Proceedings of the XVII EURALEX International congress. Lexicography and Linguistic Diversity. Pp: 284-290. http://euralex2016.tsu.ge/publication2016.pdf (accessed: November, 1st 2016).

Using Big Data to Ask Big Questions

Leah Weinryb Grohsgal
lgrohsgal@neh.gov
National Endowment for the Humanities
United States of America

How can you use open data to explore history?

Historic American newspapers illuminate the rich history and texture of life. They contain stories about politics, sports, music, shopping, food, health, science, movies, and everything in between. From the affairs of everyday life to major international events, newspapers enable humanists to delve into the past, providing vantage points in big cities and small towns, east and west, north and south, from various political, religious, and cultural standpoints, and multiple language and ethnic communities.

But what happens when cultural institutions make a large-scale set of historic newspaper data available? How are digital humanists able to use aggregated data from thousands of American news publications to tell stories and provide analysis, from the quotidian to the momentous? To find out, the U.S. National Endowment for the Humanities (NEH) asked the public "how can you use open data to explore history?" in the Chronicling America Historic American Newspapers Data Challenge. This paper describes the challenge and its results, which serve as powerful reminders of the possibilities for matching open data collections with the tools and questions of the digital humanities.

The nationwide competition, led by the NEH Division of Preservation & Access, interrogated the possibilities for using the data in *Chronicling America*, a digital repository of historic United States newspapers. *Chronicling America* is an open access, searchable database of historic U.S. newspapers, with a newly expanded date range from 1690 to 1963 (the date range at the time of the challenge was 1836 to 1922). The database and web site are produced by the National Digital Newspaper Program, a long-term partnership between the NEH and the Library of Congress (LC). *Chronicling America* includes millions of pages of digitized newspapers—nearly 12 million at the time of this writing, with more being added all the time—and descriptive information contributed by states and territories across the country. The LC supports the long-term management of the collection, including providing open access to the data through a well-documented API to enable exploration of the collection in a variety of ways beyond the site's web interface.

To spur use of the API and collection, the Chronicling America Historic American Newspapers Data Challenge encouraged researchers to create digital humanities projects, big or small, using the newspaper data. The results demonstrate exciting possibilities for connecting the creators of digital collections and the humanities research communities using them. The contest urged entrants to use the data to show trends, insights, themes, or stories from history using newspaper data. The challenge's parameters were broad, encouraging entrants to be creative in thinking about the humanities questions they find most compelling, and how they might approach them using open humanities data from newspapers.

The NEH awarded six prizes for the Chronicling America Historic American Newspapers Data Challenge in 2016. Far from being theoretical or speculative, the contest and the winning projects highlight the practical ways in which this data can be used to explore a variety of humanities themes. In brief, the winning projects were:

America's Public Bible: Bible Quotations in U.S. Newspapers, by Lincoln Mullen. The site tracks Biblical quotations in American newspapers to show how the Bible was used for cultural, social, religious, and political purposes, and how it was a contested yet common text.

American Lynching: Uncovering a Cultural Narrative, by Andrew Bales. This site explores America's long, dark history of lynching and the role of newspapers as both catalysts for killings and platforms for reform.

Historical Agricultural News, by Amy Giroux, Marcy Galbreath, and Nathan Giroux. The site is a tool for exploring information on farming organizations, technologies, and practices, as a window into social, economic, and political history.

Chronicling Hoosier, by Kristi Palmer, Caitlin Pollock, and Ted Polley. This site tracks the origins of the word "Hoosier," its geographic distribution, and its positive and negative connotations over time.

USNewsMap.com, by Claudio Saunt and Trevor Goodyear. This site allows users to discover patterns, explore regions, and investigate how words, terms, and news spread.

Digital APUSH, by Ray Palin and the A.P. U.S. History Students at Sunapee High School. This class used word frequency analysis to discover patterns in news coverage of several major issues such as secession, the KKK, and Plessy v. Ferguson.

This paper illustrates the potential impact of humanities collections such as *Chronicling America* when their data are made freely and easily available. It describes the data available in this huge data repository, the mechanisms researchers and students can use to access it, and some of the challenges inherent in its large span. The paper addresses some of the technical specifications and program guidelines for the National Digital Newspaper Program, in which metadata standards for both access and preservation were primary concerns for the NEH and the Library of Congress in creating and maintaining the program. It also explains how the program's decade-old community has cultivated shared practices and specifications that contribute both to the longevity of this dataset and to other newspaper digitization efforts across the country.

Then, the paper gives information about the winning entries in the first Chronicling America Historic American Newspapers Data Challenge in 2016, which touched on a variety of important humanities themes like religion, race, literature, violence, agriculture, law, and geography. It explains how winners used cutting-edge technology to produce maps, visualizations, tools, and data mashups. Winners reported their initial questions, the importance of this data in answering them, their methods, and their future plans for the projects they built. This paper highlights their processes and the variety of results they produced.

Finally, the paper discusses broader lessons for humanities users of "big data" in *Chronicling America*. It shows the value of a contest in publicizing data in the humanities and its uses. It showcases different modes of collaboration among humanities researchers, libraries, information technology professionals, and other partners in the research endeavor. The paper also explores some of the challenges for humanities scholars working with large datasets, including representation, bias, categorization, and documentation. The intellectual work of the digital humanities projects described here involves the same problem identifying and research that humanists have always pursued, but the methods and investigation present new questions and opportunities. The paper suggests ways of maximizing the benefits of large datasets such as *Chronicling America* to the humanities.

Towards a New Approach in the Study of Ancient Greek Music: The Virtual Reconstruction of an Aulos “Early Type” from Sicily

Angela Bellia
angela.bellia@unibo.it
University of Bologna, Italy

The ancient site of Selinunte (Sicily) is recognised today as one of the most important archaeological sites of the Greek period in Italy. From its foundation as a colony around the second half of the VII c. through to the middle of the III c. BCE, Selinunte enjoyed a prosperous existence as reflected in its notable sanctuaries, and temples. In 2006, the Institute of Fine Arts at New York University began a research project on the acropolis under the direction of Prof. C. Marconi and in collaboration with the Archaeological Park of Selinunte. The project consists of a new, systematic and interdisciplinary study of the archaeology and architecture of the main urban sanctuary, beginning with its southern sector. In the years between 2006 and 2012, the survey also consisted of the investigation of the ‘South Building’. This structure is notable for its prominent position within the sacred space. From the time of its discovery (1876), its scale has attracted scholarly attention, and the ‘South Building’ has played a significant role in discussions about the spatial articulation and cults of the main urban sanctuary. The investigation included a systematic programme of documenting the buildings in the area and its digital reconstruction.

Several elements suggest the identification of the ‘South Building’ as an impressive theatral viewing area with particular acoustic qualities (Marconi and Scahill, 2015). This building belong to a group of theatral structures found in various regions of the Greek world. Many of these structures were not proper “theatres”, but rather primitive rows of seats (meaning non-canonical theatres, with linear and non-circular *theatra* and/or *orchestra*): they existed as “a place from where one could watch”, which is in fact the original meaning of the word *theatron*.

The quality of the stones was carefully selected in relation to their placement in the cultic theatres, based on structural, aesthetic, and, acoustic considerations. The acoustics of the linear *theatra* in the Greek world have never been analysed (Blesser and Salter, 2011): no study has focused on the acoustics of these theatres in order to understand how and why these spaces were chosen for performance. It is a category of buildings brought to the attention of scholarship by Anti (1947), but first investigated in relation to religious contexts by Nielsen (2002) in her study on cultic theatres in the ancient world; by Marconi (2013) in his study on the theatral structure of Selinunte; and by Hollinshead (2015) in her study on the steps as components of monumental construction at Greek sites as early as the VI c. BCE.

At Selinunte, the cultic theatre was built to accommodate spectators of performances associated primarily with Temple R, probably a temple of Demeter. One of the main striking finds among the votive depositions was the discovery of two parts of a bone *aulos*, which can be dated to 570 BCE (Marconi, 2014). This discovery is very significant, particularly with regard to the performance associated with the activity of Temple R (Marconi, 2013). The discovery shows the importance of music in this context which already existed in the Early Archaic period, that is, since its foundation.

In 2013, the Marie Curie Actions programme (IOF) funded the TELESTES project (622974): the main aim of this project was the reconstruction of musical development at Selinunte on the basis of material culture and written sources. It also included the study of the *aulos* from Temple R using a CT scan, an improved method for the visualisation and analysis of the instruments (Bellia, 2015). However, this survey still lacked an appropriate connection between the acoustics of the cultic theatre and musical activity related to the building. Thanks to the present project, auralisation techniques are used to explore the spatial dimension of sound in this cultic theatre, establishing a relationship between the spatial configuration of this structure and how this complements music. These data will help us to understand the aural perception of ancient peoples and the type of sound experiences they were exposed to.

This research is the first study on the virtual reconstruction of the acoustics of cultic theatres of the Greek world. Despite their relevance to the field of ancient music, no study has focused on the acoustics of these theatres in order to understand how and why these spaces were chosen for performance. Within this context, archaeoacoustics is being used as a new method for the analysis of historical heritage, enabling the evaluation of the sound quality of a space (Scarre and Graeme, 2006) by using auralisation techniques which allow cognitive and physical elements to be reproduced and combined (Eneix, 2014).

This project offers an innovative research method in the study of ancient Greek music, not only in its contextualisation of archaeological evidence, but also by making connections between digital and acoustic techniques. It is also hoped that the results will provide some foundations from which to create interpretative reconstructions of what the cultic theatres might have sounded like, using digital and acoustic technology. Moreover, this study will contribute towards overcoming the traditional methods in measuring ancient instruments, opening up a new disciplinary framework for the *interaction between 3D reconstruction of instruments, their respective sounds*, and the spaces of musical performances.

In order to analyse the acoustic characteristics of the cultic theatre of Selinunte, an acoustic survey will be carried out using an impulsive sound source located in two positions on the steps, and in the "orchestra". Measurements will be recorded using an impulsive sound source. This process will make it possible to obtain a virtual acoustic reconstruction of this space, including the incorporation of acoustic characteristics as another important aspect of its intangible heritage. The software used will be CATT-Acoustic, v. 9.0 c.

The digitisation process of the *aulos* is divided into two main tasks, namely the 3D scanning phase and the post-processing phase. The tools we plan to develop are divided into those involving the use of computational methods for processing the 3D models, and those involving the development of interactive tools aimed at engaging in the exploration of the instruments. The software used will be GEOMAGIC DESIGN X, with AVIZO, v. 9.0. Following the digital model will translated into artificial copies, using polymer as a material. We aim to obtain and to assess the auralisation of the acoustical properties of these 3D models.

In addition, the work will address the study of written sources as well as visual and archaeological documentation related to music performed in the cultic theatres of the ancient Greek world from the Archaic through to the Classical periods. The study will be conducted in order to understand the reasons that led ancient cultures to create these spaces, as well as reconstruct how they experienced them.

Expected results

Firstly, this research will provide the first acoustic model for the study of acoustical properties of cultic theatres in the ancient Greek world. The study will assess and recover the acoustics of the Selinunte theatre.

Secondly, the research will develop specific tools suitable for processing the resulting 3D models. It is also hoped that the results will provide some foundations from which to create experimental interpretative 3D reconstructions integrating acoustic models. The results will establish a new framework, which future researchers can use to advance their knowledge of the application of 3D technology for the documentation of instruments.

Finally, this project will offer an innovative research method in the study of ancient Greek music. This research aims to create a field of comparative studies of archaeomusicological research.

In conclusion, this research will develop a new theoretical basis, which will contribute to the establishment of a methodology at the crossroads of archaeomusicology, architecture, and acoustics and digital technologies. In addition, this study is part of a programme intended to valorise ancient cultural and musical heritage in the Mediterranean with cross-disciplinary approaches to human culture and technology, in order to unveil new meanings and create new research fields within the digital humanities and heritage science.

Bibliography

Anti, C. (1947). *Teatri greci arcaici*. Padova: Le Tre Venezie.

Bellia, A. (2015). "*The Virtual Reconstruction of an Ancient Musical Instrument*." In Guidi, G., Scopigno, R., Torres, J.C., Graf, H. (eds), *2015 Digital Heritage*, I, Proceedings of the International Congress DigitalHeritage 2015 (Granada (Spain), 28 September – 2 October 2015). IEEE, pp. 55-58.

Blesser, B.A., and Salter, L.-R. (2011). "Beyond Measurements: A Multi-disciplinary Framework for Aural Experience of Ancient Spaces." In *The Acoustics of Ancient Theatres*, Patras, September 18-21, 2011: http://www.blesser.net/downloads/Blesser-Salter%20Beyond%20Measurments.pdf;

Eneix, L.C. (2014) (ed.). *Archaeoacoustics. The Archaeology of Sound*. Myakka (Fl): OTS Foundation.

Hollinshead, M.B. (2015). *Shaping Ceremony*. Madison: The University of Wisconsis Press.

Marconi, C. (2013). "Nuovi dati sui culti del settore meridionale del grande santuario urbano", *Sicilia Antiqua*, 10: 263-271;

Marconi, C. (2014). "A New Bone Aulos from Selinus". In Bellia A. (ed.), *Musica, culti e riti dei Greci d'Occidente*. Roma-Pisa: Fabrizio Serra, pp. 105-116.

Marconi, C., and Scahill, D. (2015). "The "South Building" in the Main Urban Sanctuary of Selinus: A Theatral Structure?." In Frederiksen, R., Gebhard, E.R., and Sokolicek, A. (eds), *The Architecture of the Ancient Greek Theatre*. Aarhus: University Press, pp. 281-294.

Nielsen, I. (2002), *Cultic Theatres and Ritual Drama*. Aarhus: University Press.

Scarre, C., Graeme, L. (2006) (eds.), *Archaeoacoustics*. Cambridge: McDonald Institute for Archaeological Research.

Suárez, R., Alonso, A., and Sendra, J.J. (2016), "Archaeoacoustics of Intangible Cultural Heritage", *Journal of Cultural Heritage*, 19: 567-572.

Optical Character Recognition with a Neural Network Model for Printed Coptic Texts

Kirill Bulert
kirill.bulert@stud.uni-goettingen.de
eTRAP Research Group
University of Göttingen, Germany

So Miyagawa
so.miyagawa@mail.uni-goettingen.de
University of Göttingen, Germany

Marco Büchler
mbuechler@etrap.eu
eTRAP Research Group
University of Göttingen, Germany

Introduction

Optical character recognition (OCR) is the process of extracting text from images. The final results are machine readable versions of the original images. Nowadays every modern scanner comes with some kind of OCR, but the results may not be satisfying when the OCR is applied to historical texts, that

1. do not use standard fonts,
2. are not printed by a machine,
3. have varying paper and font quality.

Furthermore, historical texts are not passed down through the centuries in their entirety but rather contain lacunae and fragmentary words. This makes automatic post-correction more difficult on historical texts than on modern ones.

We used two tools to create language- and even document- specific recognition patterns (or so-called models) to recognize printed Coptic texts. Coptic is the last stage of the pre-Arabic, indigenous Egyptian language. It was used to create a rich and unique body of literature: monastic, "Gnostic," Manichaean, magical and medical texts, hagiographies, and biblical and patristic translations. We found that Coptic texts have properties which make them excellent candidates for reading by computers. The characters can easily be distinguished due to their limited number and the fact that almost all the hand-written texts exhibit characters with highly consistent forms.

Related Work

The process of digitizing historical documents can be split up into at least three major steps: (1) pre-processing, (2) text prediction (OCR), and (3) post-processing or correction.

Although many works already tackled subproblems (He et al, 2005; Gupta et al, 2007; Kluzner et al, 2009), Springman et al.(2014) presented the first complete approach containing all major steps for historical Greek and Latin books.

The first OCR results for printed Coptic texts were achieved by Mekhaiel (see Moheb's Coptic Pages) by using Tesseract to create a model for Coptic texts. Tesseract assumes that the image was printed with a standardized font. Although it can be trained to use many different fonts, creating a general model that would satisfy scholars is not feasible. In the end, this model is sufficient for pure printed Coptic texts, but creates a lot of noise for texts with mixed languages or annotations. Such drawbacks can be easily overcome by checking against a dictionary, but historical languages often do not have a dictionary that could be considered complete, and the texts might only be fragments that require further analysis.

The recognition itself is performed by either Ocropy (Breuel, 2008) or Tesseract. Potentially, all character-based texts can be recognized. However, even though Mekhaiel provided a Coptic model for Tesseract, we were never able to achieve satisfying results on images which were not pre-processed.

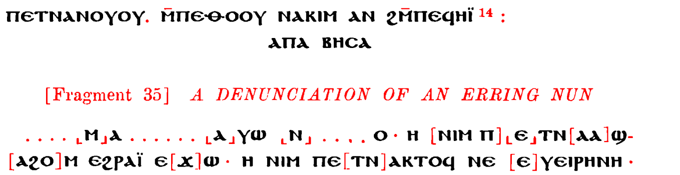
Data Used

For training and testing, an expert on Coptic created a clean version and transcription of Kuhn's 1956 edition "Letters and sermons of Besa." This will also be made available to the interested public.

Besa is a fifth-century abbot of a monastery in Upper Egypt and Coptic writer, whose literary legacy consists mainly of letters to monks and nuns on questions of monastic life and discipline.

Simplified pages were created to find the limits of the trained models with optimal input data. Since creating simplified pages consumes a lot of time, we consider this task as impractical for real use scenarios. Nevertheless, the results on these simplified pages show the best possible prediction.

In Fig. 1 all characters and symbols that are going to be removed are marked red. The resulting simplified image can be seen in Fig. 2. By procedure, adjacent characters that are supposed to form one word are cut apart by gaps. Those gaps are going to be predicted differently by the two OCR engines.

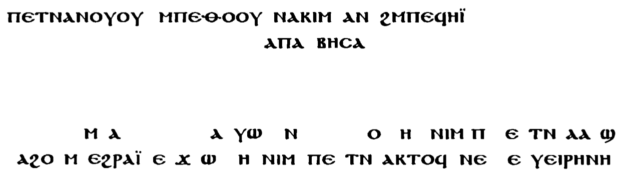


ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ. ⲙ̄ⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙ̄ⲡⲉϥⲏⲓ̈ [14] :
ⲁⲡⲁ ⲃⲏⲥⲁ

[Fragment 35] *A DENUNCIATION OF AN ERRING NUN*

. . . . ⸤ⲙ⸥ⲁ . . . . . . ⸤ⲁ⸥ⲩⲱ ⸤ⲛ⸥ . . . . ⲟ · ⲏ [ⲛⲓⲙ ⲡ]⸤ⲉ⸥ⲧⲛ[ⲁⲁ]ϣ-
[ⲁϩⲟ]ⲙ ⲉϩⲣⲁⲓ̈ ⲉ[ϫ]ⲱ · ⲏ ⲛⲓⲙ ⲡⲉ[ⲧⲛ]ⲁⲕⲧⲟϥ ⲛⲉ [ⲉ]ⲩⲉⲓⲣⲏⲛⲏ ·

Fig.1, Original Image (excerpt), red elements are missing in the simplified version



ⲡⲉⲧⲛⲁⲛⲟⲩⲟⲩ ⲙⲡⲉⲑⲟⲟⲩ ⲛⲁⲕⲓⲙ ⲁⲛ ϩⲙⲡⲉϥⲏⲓ̈
ⲁⲡⲁ ⲃⲏⲥⲁ

ⲙ ⲁ ⲁ ⲩⲱ ⲛ ⲟ ⲏ ⲛⲓⲙ ⲡ ⲉ ⲧⲛ ⲁⲁ ϣ
ⲁϩⲟ ⲙ ⲉϩⲣⲁⲓ̈ ⲉ ϫ ⲱ ⲏ ⲛⲓⲙ ⲡⲉ ⲧⲛ ⲁⲕⲧⲟϥ ⲛⲉ ⲉ ⲩⲉⲓⲣⲏⲛⲏ

Fig. 2, Simplified version (excerpt)

Methodology

There are two methods to train for Coptic texts:

(i) Tesseract needs a font as the baseline and matches the found letters against this font. This can be highly convenient since fonts do not show many variations within a single document. Additional fonts can be incorporated into the model with the drawback that the prediction requires more computational time. So far, we have used Mekhaiel's original model, and we are currently experimenting by adding document-specific characters to increase the accuracy of a single document.

(ii) Ocropy, on the other hand, does not require a font. For training, it requires only a partial transcription: the ground truth. This transcription is used to train a neural network that can recognize the characters. Ocropy's drawback is that the ground truth cannot just be the alphabet but requires multiple pages of transcribed text with a representative letter frequency. Ocropy's training process is measured in iterations. Springmann proposed working with at least 30,000 iterations (a comment made by Springmann in a private conversation, based upon his own experience).

For this contribution, we created an Ocropy model with a training set containing approximately 5,000 characters. This set includes superlinear strokes, braces and foreign characters which are not part of the Coptic alphabet.
Multilingual documents and documents containing foreign characters are considered complex. Stains on the document, bad image quality, and annotations like line numbers increase the complexity of documents as well. We, therefore, created special pages with reduced complexity. Our original pages were stripped offline numbering and footnote annotations. In the "clean" version, all foreign characters, punctuations and annotations inside the text were removed, leaving us with a pure Coptic text. We further stripped all clean versions of superlinear strokes, giving us the simplified version.

For testing, the selected pages were transcribed with corresponding 'original', 'clean' and 'clean without stroke or simplified' ground truths. All results were compared with 'Ocreval' (Baumann 2014)[9] against the ground truth.

Results

Prediction

Mekhaiel's original Tesseract model produced the best results on simplified pages with an accuracy of ~95%, while our Ocropy model performed better on the more complex pages. On the other hand, the Tesseract tends to produce predictable errors. Character ϣ will, for example, always be recognised as □; while, Ocropy produces unpredictable errors. Although our Ocropy model is less accurate on simplified pages, it surpasses Tesseract on noisier pages.

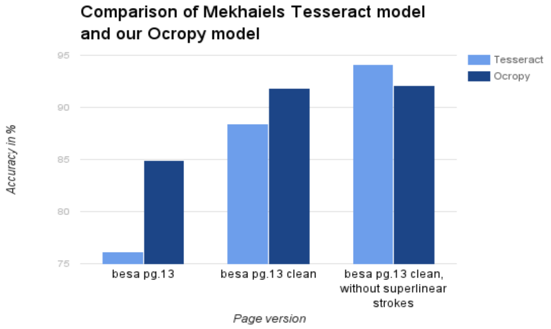


Fig. 3, OCR accuracy on different complexity levels

Costs

We measured that a skilled person needs roughly 10 minutes for manual transcription and 5 additional minutes for proofreading per page. Ocropy's models are built on top of transcribed images. Therefore, an initial ground truth is always required. Training with Ocropy does not require further human interaction but consumes up to two days of CPU power (Core i3/5 2.4GHz/3.2GHz, 8GB RAM, SSD), training cannot be run in parallel. Tesseract's training process, on the other hand, depends on the font extraction. We do not have enough data to estimate the time required to extract a font from an image. Both predictions still have to be checked manually, which can take up to 5 minutes. With clean pages and reduced proofreading time per page, Fig. 4 shows an optimal OCR workload reduction (red lines) in comparison to manual transcription (yellow line). A more realistic scenario is mentioned in the discussion.

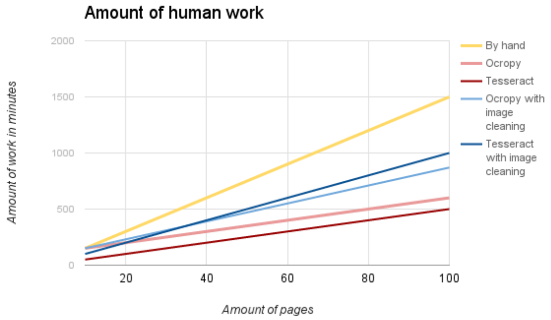


Fig. 4, workload comparison

Discussion

Our result shows that Tesseract outperforms Ocropy on simplified pages in terms of accuracy and amount of human work. Unfortunately, in a realistic scenario, the pictures will always contain some of the previously described complexities. Pre-processing of the data is, therefore, essential to obtain good results. In Figure 4, we also computed a more realistic scenario (blue lines) with a higher workload on pre-processing for Tesseract. It shows that creating an Ocropy model pays off for larger and more complex document sets.

Tesseract's overall acceptable performance is based on the fact that no model has to be trained. As creating and testing a model can consume more time than manual transcription and proofreading, the creation of clean images might still be less efficient than the manual approach even if a model can be reused.

As long as cleaned images images are one of the desired results, our works shows that the workload can be reduced by half. This applies especially to Ocropy, since ground truth creation and training fit into the normal transcription workflow.

Unicode ambiguities, which unfortunately result in encoding differences, require normalization and filtering. Otherwise, these encoding differences, which would not be seen as errors by humans, will be counted. Due to the same ambiguities, it is easy to mix characters from different code pages, especially on multilingual texts and text markings. It is, therefore, recommended that one use only corresponding code pages, especially with multilingual models. Tests with models containing multilingual fonts will be considered in further studies.

Conclusion

OCR of historical documents continues to be a hard problem, but we showed that utilizing OCR for the transcription of Coptic texts can reduce the overall workload. Since even the simplest images could not be recognized with 100% accuracy, further gains can only be achieved by better pre- and post-processing techniques.

A bigger workload reduction can be achieved by model reuse. However, no Coptic OCR models have been published besides Mekhaiel's. Therefore, we highly recommend publishing models alongside the transcription and suggest that it is possible to predict almost all well-preserved texts.

Also, although our model was able to partially predict multilingual texts, further studies are required. Multilingual texts require a specialized training process to compensate for the small numbers of foreign words.

Bibliography

He, J., Do, Q. D. M., Downton, A. C., and Kim, J. H. (2005) "A comparison of binarization methods for historical archive documents," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, p. 538–542 Vol. 1.

Gupta, M. R., Jacobson, N. P., and Garcia, E. K. (2007). "{OCR} binarization and image pre-processing for searching historical documents," *Pattern Recognit.*, vol. 40, no. 2, pp. 389–397.

Kluzner, V., Tzadok, A., Shimony, Y., Walach, E., and Antonacopoulos, A. (2009) "Word-Based Adaptive OCR for Historical Books," in *2009 10th International Conference on Document Analysis and Recognition*, pp. 501–505.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., and Fink, F. (2014) "OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 71–75.

Mekhaiel, M. S. (n.d.) "Moheb's Coptic Pages." [Online]. Available: http://www.moheb.de/. [Accessed: 01-Nov-2016].

"Tesseract OCR." [Online]. Available: https://github.com/tesseract-ocr. [Accessed: 01-Nov-2016].

"Ocropy." [Online]. Available: https://github.com/tmbdev/ocropy. [Accessed: 13-Dec-2016].

Breuel, T. M. (2008) "The OCRopus open source OCR system," *Proc. SPIE 6815, Doc. Recognit. Retr. XV*, 2008.

Baumann, R. (2014) "OCR Evaluation Tools." [Online]. Available: https://github.com/ryanfb/ancientgreekocr-ocr-evaluation-tools. [Accessed: 01-Nov-2016].

Distinguishing Newspaper Genres. Exploring Automated Classification of Journalism's Modes of Expression

Frank Harbers
f.harbers@rug.nl
University of Groningen, the Netherlands

Juliette Lonij
juliette.lonij@kb.nl
National Library of the Netherlands, the Netherlands

Introduction

This paper examines the opportunities, approaches and issues of automatically classifying historical newspaper articles from the Netherlands for 'genre' as an expression of the historically and culturally determined conception of journalism. Genre is defined as "language use in a conventionalized communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms" (Handford 2010). As Barnhurst and Nerone (2001) argue: "The form includes the way the medium imagines itself to be and to act. In its physical arrangement, structure, and format, a newspaper reiterates an ideal for itself." Examining the generic form of newspaper articles form a historical perspective therefore sheds an interesting light on the way newspaper journalism has developed.

The digital era, which is typified as the 'age of abundance' of historical newspaper material, poses new challenges to historical research. Historical approaches to selecting and analyzing newspapers, rooted in the assumption of a scarcity of available material, had to be replaced with social scientific methods, such as quantitative content analysis (Nicholson 2013; Broersma 2009). Yet, manual quantitative content analyses are still highly time consuming and therefore expensive. Moreover, even then the size of the material that can be covered represents only a small part of the amount of material available (Harbers 2014). Automated forms of (content) analysis have the potential to alleviate or even solve this issue. As such, automatic forms of content analysis would be highly suitable for longitudinal and also comparative historical research into the development of newspapers, which particularly grapples with the overabundance of available research material (Broersma 2009).

However, although these approaches have a great appeal to researchers (Allen, Waldstein & Zhu 2008; Grimmer & Stewart 2013), research in this vein is mostly done in information science and linguistics. It seldom has a press historical perspective (Broersma 2009). Moreover, the emphasis has mostly been on topical modeling (Lee & Myaeng 2002), whereas attention for automatic classification of style and genre is scarce. Rather than determining what subjects and themes are being discussed, this project aims to examine genre as a modes of expression of newspapers, shedding light on the discursive context (Handford 2010). This is a particularly difficult task as genres are dynamic and can change or fade away over time while new ones can emerge. Moreover, genres are ideal-typical discursive constructs, which means the textual manifestations do not always match the characteristics of these constructs perfectly, nor can they always be clearly delineated from other genres.

The research question therefore is:

To what extent and how can historical newspaper articles be automatically classified for genre?

To examine this question, we have designed a research project that builds on an existing set of metadata describing several textual characteristics, such as genre, of a large sample of historical newspaper articles. This dataset was the result of a large-scale research project into the historical development of European newspapers with the title 'Reporting at the boundaries of the public sphere. Form, Style and Strategy of European Journalism, 1880-2005'. The set of metadata is derived from a manual content analysis that coded for genre based on a detailed rule-based coding manual. The metadata set relates to a corpus of approximately 33.000 Dutch newspaper articles from three types of Dutch newspapers, divided over the sample years 1885, 1905, 1925, 1965, 1985 and 2005 (Harbers, 2014). This set of metadata thus provides us with a number of labeled example articles that can be used to train and formally evaluate a classifier that is able to automatically predict the genre of additional samples of historical newspaper articles. In order to so the metadata **first** has to be connected to the full text of the corresponding digitized articles in the Dutch newspaper repository of the National Library (KB). The **second** step is to use this enriched dataset to train a classifier that can classify historical newspaper articles automatically.

This paper first shows a way to connect the metadata to the digitized historical newspaper articles (**Phase 1**). Ultimately, it offers an outline of a concrete machine learning approach, applying linear and non-linear classifiers, to predict the genre of a newspaper article. As a part of this, the paper discusses the different tools we have tried out and the problems we have encountered in the process (**Phase 2**). Specifically, the paper reflects on the way the rule-based approach to determining genre in the manual content analysis relates to the training of an automatic classifier based on machine learning techniques.

In order to link the articles described in the existing metadata set to the corresponding KB data, we first created a number of rules to find the most promising candidate links for each item in the original data set, based on the position of the article on the page, its size, and the presence of images and quotes. Since these rule alone turned out not to be sufficiently accurate, a simple classifier was trained to select the best link from the candidate set, if any, based on features such as the difference in size and number of images present between the article as described in the original data set and the article as found in the KB repository, as well as author mentions and subject matter, amongst other features. By only accepting links predicted by this classifier with a relatively high confidence value approximately 50% of all articles could be automatically linked, with an error rate of 0.5%. Some genres were underrepresented in the automatically linked set and were manually expanded upon.

The data set resulting from phase 1 was then used to create a training and test set for the actual genre classifier. This, again, involved taking several steps: first, the OCR for the articles in the set was retrieved and cleaned of quoted text by means of regular expressions. Next, the remaining text was pre-processed with the Natural Language Processing software package Frog, performing tasks such as segmentation, tokenization, and part-of-speech tagging. From the resulting annotated text, the relevant features for each article were calculated, such as number of words, number of sentences, number of direct quotes removed, and, most importantly, number of adjectives and various types of pronouns found in the text. These features, together with the existing genre labels were finally used to train a Support Vector Machine to choose one of eight possible genres for each article, ranging from news report and interview to news analysis and opinion article.

The initial results we obtained with this classifier were quite promising. Our first attempt resulted in an accuracy score of 58%, with the default accuracy by predicting the majority class or genre being 46%. The confusion matrix made from the predictions provides important information about which genres are most difficult to predict and what mistakes are most commonly made. This information can help us to fine-tune the existing features. Moreover, we have not yet implemented all the textual features that typify the different genres, such as named entities occurring in the text, sentiment, or self-classification to name only a few. Finally, we will compare the results of different algorithms, including deep learning approaches. Based on this, we expect to be able to improve these initial results in the coming months and present our final results in more detail at the conference.

Bibliography

Allen, R.B., Waldstein, I. & Zhu, W. (2008). 'Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres'. In: Buchanan, G., Masoodian, M.& Cunningham, S.J. (eds.), *Digital Libraries: Universal and Ubiquitous Access to Information*. New York: Springer

Barnhust, K. & Nerone, J. (2001). *The Form of News. A History*. New York: Guildford Press

Broersma, M. (2009). 'Nooit meer bladeren. Digitale krantenarchieven als bron'. In: *Tijdschrift voor Mediageschiedenis* 14(2): 29-55

Grimmer, J. & Stewart, B.M. (2013). 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. In: *Political Analysis* 21(3): 267-297.

Handford, M. (2010). 'What can a corpus tell us about specialist genres'. In: 'o Keeffe, A. & McCarthy, M. (eds.), *The Routledge Handbook for Corpus Linguistics*. New York: Routledge.

Harbers, F. (2014). *Between Personal Experience and Detached Information. The Development of Reporting and the Reportage in Great Britain, the Netherlands and France, 1880-2005*. PhD University of Groningen

Lee, Y. & Myaeng, S.H. (2002). 'Text Genre Classification with Genre-Revealing and Subject-Revealing Features'. SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. 145-150.

Nicholson, B. (2013). 'The Digital Turn'. In: *Media History* 19(1): 59-73.

Genetic Editions to the Extreme? Conserving Historical Text Generators

Claus-Michael Schlesinger
claus-michael.schlesinger@ilw.uni-stuttgart.de
Universität Stuttgart, Germany

Short Description

My talk will focus on the problems and possibilities of conserving historic text generators, 'historic' meaning roughly 1950-1970. For this purpose, I will develop the scientific interests in these generators either from the perspective of literary history and from the perspective of the history of knowledge, since many of the text generators in question were connected explicitly to scientific interests.

A system for conserving and editing historic text generators, that will enable researchers from both of these fields to access historic text generators in order to study their esthetics and functioning needs, in orderto take into account not only the generated texts, but also the generating texts, meaning software and prerequisites such as flowcharts. In my talk, I will focus on the materiality and the potentiality of the historic text generators in order to propose a platform solution that should enable scholars to edit (and publish) historic textgenerators.

Full Abstract

In their manifesto Zur Lage (Bense and Döhl (1964)) (State of Things) Max Bense and Reinhard Döhl name several tendencies of contemporary literature. The sixth and last point of their list calls on a cybernetic and material poetry ("kybernetische und materiale Poesie"). Historically and biographically, Bense and Döhl were placed at one of the centers of concrete poetry and concrete art (Stuttgart, Germany), and their manifesto reflects the artistic and academic developments made in these areas. But already in 1964, when Bense and Döhl published their manifesto, experiments had taken place that had taken the notion of a cybernetic poetry quite literally, using computers and software to generate literary texts. (Strachey Okt. 1954; Lutz 1959; Levin 1963; Gunzenhäuser 2004; Link 2004) The automatic generation of texts by way of combining syntax, vocabulary and a (pseudo) random generator that serves to fill the syntax positions with proper words from the vocabulary has a history dating back to the 17th century. However, it was the use of computers that prompted artists and researchers to expand their experiments and develop complex setups applying hundreds of syntactical structures and bigger vocabularies (as did, for example, Stickel [1967]).

The research on early text generators has focused on biographical data and the literary history of automatic text generation (Bülow 2007), the autoreflexivity of digital texts (Cramer 2003) and the genealogy of text generating systems (Link 2004). A literary and cultural history of text generators needs to take into account not only the texts that were produced, but also the modes and methods of production. In my talk, I will argue that in order to understand the history of automatic text generation, we need to join these different perspectives and research either the generated texts and the modes and methods of text generation.

Now the first problem when dealing with text generators (and the texts generated with them historically) is the availability of the material. Only a small fraction of historically generated texts has been published, and in most cases the documentation of the conceptual and technical setup is only partly available through publications, not to mention the actual code that was used to produce the text in the first place. This is due to the fact, that literary scholars have tended to ignore texts of this kind – that is, experimental text which barely fulfills any definition of poetry – and are only beginning to recognize them as a certain type of literature in its own kind. The second reason for a newly developed recognition for early text generation experiments is the fact that automatic text generation has become ubiquitous in our time, from simple twitter bots that employ the same methods as the early text generators (meaning simple syntactical structures filled with randomly chosen words), via the modern ELIZAs of contemporary electronic assistant systems, to complex machine learning algorithms that synthesize Shakespearean plays. (Goodwin 2016)

So in order to further research on early text generators, the multiple dimensions of text generators will have to be made available to researchers from all fields concerned with the history of automatic text generation. Since I am, by profession, a philologist, my approach is limited to the document side of things and does not consider hardware conservation or emulation.

Documents connected to text generators comprise flowcharts, punched tape, source code data on magnetic memory, print-outs, project documentation in various paper formats. For a digital representation, these materials can be digitized either in graphical or in textual format. Graphical representations can be annotated with their respective metadata describing their material, function, authorship etc. As for the texts contained in a text generator as a historic object, we are dealing with three different kinds of texts:

- texts that have been generated
- texts that have been used to generate texts (mainly software)
- texts that could have been or can potentially be generated (the full outcome of the underlying algorithm)

Texts in groups 1 and 2 are connected in a genetic way. If we take the Stochastic Texts by Theo Lutz and Rul Gunzenhäuser (Lutz 1959) as an example, we can see that there is a published version of the generated texts and a raw version (published in Büscher, 2004). The differences between these versions are significant. From a text-genetic perspective, these differences lead to the production of the texts or, in other words, the execution of the underlying program. The execution of the program leads to the implementation of the program or the source code, the source code leads to the conceptual implementation of the algorithm (e.g. flowcharts) and thus to the abstract underlying algorithm.

Thus, a genetic perspective on text generators leads not only to the raw version of the output of the computer, but to the implementation of the algorithm in source code as well. The title of my talk searches to reflect this notion of text genesis, because the link between source code and output is not a certain similarity between two texts, but a functional, algorithmic connection. I think that problematizing this kind of functional genetic connection can also be fruitful for philology dealing with texts where computers don't generate all of the text, but help in generating certain features of the text. In order to describe and show this kind of connection in different text generators, it will be necessary to develop standardized notions for annotation schemas (Currently, the annotation of source code is not part of the TEI Guidelines (TEI-Consortium and Lou Burnard 2014)

Finally, studying the poetics of a text generator might also mean to study more generated text than is (historically) available. A representation of a text generator thus not only should comprise the source code itself, but also the possibility to run the algorithm in order to generate some of the potential texts. The best solution would be, of course, to run the code on the historic machines it was developed for. Not only would this repeat the original experiments, it would also hopefully reproduce the glitches that can be seen in the raw output that has been preserved and archived. (Büscher 2004) However, such a procedure would be way too costly. An easier way is to implement the reconstructed algorithm in a modern programming language and to offer researchers the possibility to run the code without being forced to undertake major retrocomputing tasks just in order to gather some more textual outcome.

Since this is a work in progress, I am confident that by August 2017 I will be able to outline not only the problems, but also some solutions for implementing a platform and database to preserve and present historic text generators.

Bibliography

Büscher, B., von Herrmann, H.-C., Hoffmann, C. (2004). *Ästhetik Als Programm. Max Bense/Daten Und Streuungen.* Berlin: diaphanes.

Bense, M., and Döhl, R. (1964). *Zur Lage.*

Bülow, R. (2007). "SINN IST FERN - Wie Die Computer Dichten Lernten." In *Ex Machina - Frühe Computergrafik Bis 1979: Die Sammlungen Franke Und Weitere Stiftungen in Der Kunsthalle Bremen*, 134–72. München: Deutscher Kunstverlag.

Cramer, F. (2003). Exe.cut[up]able Statements. Das Drängen Des Codes an Die Nutzeroberflächen. Berlin.

Goodwin, R. (2016). "Adventures in Narrated Reality. New Forms & Interfaces for Written Language, Enabled by Machine Intelligence." March. https:// medium.com/artists-and-machine-intelligence/adventures-in-narrated-reality-6516ff395ba3#.5ko9ojqfq.

Gunzenhäuser, R. (2004). "Zur Synthese von Texten Mit Hilfe Programmges- teuerter Ziffernrechenanlagen." In *Ästhetik Als*

Programm. Max Bense/Daten Und Streuungen, edited by Barbara Büscher and Hans-Christian von Herrmann Christoph Hoffmann, 170–83. Berlin: diaphanes.

Levin, S. R. (1963). “On Automatic Production of Poetic Sequences.” *The University of Texas Studies in Literature and Language* V: 138–46.

Link,D. (2004). *Poesiemaschinen-Maschinenpoesie.* Berlin: Diss. Humboldt- Universität.

Lutz, T. (1959). “Stochastische Texte.” *Augenblick* 4 (1): 3–9.

Stickel, G. (1967). “Monte-Carlo-Texte. Automatische Manipulation von Sprachlichen Einheiten.” In *Kunst Aus Dem Computer*, edited by William E. Simmat, 53–57. Stuttgart: nadolski.

Strachey, C. O. (1954). “The ’Thinking’ Machine.” *Encounter* 3 (4): 25–31.

TEI-Consortium, Bauman, S., Burnard, L. (2014). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 2.6.0, 20.1.2014, Revision 12802. o. O.

WeisoEvent: A Ming-Weiso Event Analytics Tool with Named Entity Markup and Spatial-Temporal Information Linking

Richard Tzong-Han Tsai
thtsai@csie.ncu.edu.tw
Center for GIS, Academia Sinica, Taiwan

Yu-Ting Lai
trulight@hotmail.com.tw
National Central University, Taiwan

Pi-Ling Pai
lingpai@gate.sinica.edu.tw
Center for GIS, Academia Sinica, Taiwan

Yu-Chun Wang
albyu35@gmail.com
Chunghwa Telecom, Taiwan

Sunny Hui-Ming Huang
shtilberg0623lg@gmail.com
Institute of History and Philology,
Academia Sinica, Taiwan

I-Chun Fan
mhfanbbc@ccvax.sinica.edu.tw
Center for GIS, Academia Sinica, Taiwan

Introduction

Weiso(衛所制), which means "guardhouse", is one of the military units of the barracks used by the Chinese dynasty Ming (1368-1644) to maintain peace throughout its empire. WeisoEvent is a web-based digital humanity research tool targeting Ming Weiso events recorded in Ming Shilu, which contains the imperial annals of the Ming emperors. WeisoEvent is composed of two parts: (1) an event type classifier that categorizes paragraphs according to their event types; (2) an analytics tool that shows (1)’s result, markups named entities, links guard mentions to Academia Sinica’s Chinese Civilization in Time and Space (CCTS) spatial-temporal platform, and provides four visualization functions. Historians can use this tool to search for specific event types and gain insight into the relationship between particular guards and those event types, not only improving the efficiency but still maintaining the quality of research.

Event type classifier

Normally, one would develop a supervised-learning-based text categorization system to classify paragraphs into different event types. This involves defining a set of categories and annotating example texts for each category. However, lacking the human resources needed for such a task, we use unsupervised text clustering, which groups paragraphs into clusters by event type, to generate categories and their corresponding paragraphs for training an automatic event classifier. Although the results are not as accurate as those of pure supervised text classification, this hybrid approach is an acceptable tradeoff.

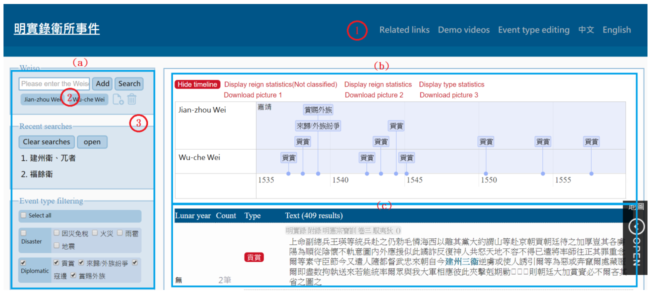
In clustering algorithms, each paragraph is represented as a vector. In previous studies, paragraphs have been represented using the vector space model (VSM), which represents each text as a feature vector of terms. However, this approach loses the ordering and ignores semantics. Yet another representation scheme inspired by word2vec is the “Paragraph Vector” proposed by (Le and Mikolov, 2014), an unsupervised framework that learns continuous distributed vectors for pieces of text. In their model, entire paragraphs are represented as vectors. The vector representation is trained to predict the words in a paragraph. More precisely, they concatenate the paragraph vector with several word vectors from a paragraph and predict the following word in the given context. Le’s Paragraph Vector model has many advantages. First, it is mostly unsupervised and works well with sparsely labeled data. Second, it is suitable for text strings of various lengths, ranging from sentences to whole documents. Finally, it can overcome many weaknesses of the bag-of-words and bag-of-n-grams models. Because it does not suffer from data scarcity and high dimensionality, it also preserves the ordering and semantic information.

In summary, we propose a classification method which is based on clustering. First, we employ a named entity (NE) recognizer to label texts. Second, we train a paragraph vector model to represent paragraphs as vectors. Third, we cluster paragraphs with length <40 characters. Finally, we use the clustering results as gold-standard categories with which to train a support-vector-machines classifier to predict other paragraphs' categories.

We compare our method with the state-of-the-art paragraph clustering method using continuous vector space representation proposed by (M. Chinea-Rios et al., 2015). They use word2vec to learn word vectors and represent each sentence by summing the vectors of the words in that sentence. Like Chinea-Rios et al., we use the k-means algorithm to cluster vectors. We set the number of clusters to 68. We refer to the evaluation measures used in (Le and Mikolov, 2014). We generate sets of three paragraphs: two with the same event type and one with a different event type. Each set is referred to as a paragraph triplet. The distance between the two vectors with the same event type should be closer than the distance between either of these two and the unrelated one. We collect 923 paragraph triplets and compute the accuracy. Our best configuration that combines word dimensions and named entity dimensions to generate paragraph vectors achieves an accuracy of 62.49%, outperforming Chinea-Rios et al.'s pure text-clustering approach (M. Chinea-Rios et al., 2015) by 24.65%.

Analytics tool interface

WeisoEvent groups paragraphs with similar subjects into clusters automatically and each cluster is named manually according to the main subject of its paragraphs. Clusters with related topics are grouped into broader categories for search convenience. For example, we group "earthquake", "conflagration" and "hailstorm" into the event category "disaster". Users can modify event category titles by clicking a button on the top-right of the webpage (see Fig. 1, [1])

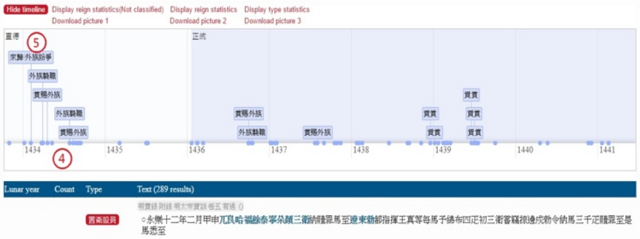


Lunar year	Count	Type	Text (409 results)
無	2筆	貢賞	明實錄 附錄 明憲宗寶訓 卷三 馭夷狄 () 上命副總兵王瑛等統兵赴之仍勅毛憐海西以離其黨大約謂山等赴京朝貢朝廷待之加厚豈其各處陽為順從陰懷不軌意圖内外應援似此譎詐反復神人共怒天地不容不得已遣將率師往正其罪重念爾等素守臣節今又遣人隨都督武忠來朝自今建州三衞逆虜或使人誘引爾等為惡或奔竄爾處藏匿爾即盡数拘執送來若能統率爾眾與我大軍相應彼此夾擊剋期勦□□□則朝廷大加賞賚必不爾吝其省之圖之

Figure 1. System interface.

Fig. 1 shows our research tool webpage, which consists of three main windows: (a) search parameters, (b) search result visualizations, and (c) search results snippets.

1. Search parameters: Users can search for one or more guards by typing the name into the search box (Fig. 1, [2]). For convenience, a user may also import a guard list by clicking the "import" icon (Fig. 1, [3]). Notably, if two guards are queried, their event timelines are displayed in parellel, like Jian-zhou (建州) and Wu-che (兀者) guards shown in window b. After at least one event category is selected, the search results are shown in windows b and c.
2. Search result visualization tools: In window b, users can select among four visualization options at the top of the frame. Option 1 hides or shows an event timeline of the search results on the page. When the event timeline is enabled, event-type labels corresponding to each retrieved paragraph are displayed chronologically on the timeline. For reference, the timeline shows the CE year at the bottom of the window (Fig. 2, [4]) and the name of era, which usually corresponds with the reigning Ming emperor, at the top (Fig. 2, [5]). When a user clicks on an event icon in the timeline, the corresponding text snippet is displayed in window c, highlighted in yellow (Fig. 2). Figure 3 takes Jian-zhou guard (建州衛) as an example to depict options 2 to 4. Option 2 is the bar chart. Each bar corresponds to a Chinese era name and represents the total number of paragraphs for the three selected event types in that era. Option 3 shows each bar sub-divided by color to show the distribution of paragraphs of each event type ("come over and pledge allegiance"/ "reward alien"/ "tribute-reward") in each era. By clicking Option 4, a pie chart shows the distribution of the three selected event types in the entire dataset. The slice for each event type is labeled with the number of paragraphs of that event type and its percentage of the total. These data visualizations offer historians a quick statistical overview of selected event types.



Lunar year	Count	Type	Text (289 results)
		置衛設員	明實錄 附錄 明太宗寶訓 卷五 宥過 () ○永樂十二年二月甲申兀良哈福餘泰寧朵顏三衛納贖罪馬至遼東勅都指揮王真等每馬予綿布四疋初三衛嘗竊掠邊戍勅令納馬三千疋贖罪至是馬悉至

Figure 2. Event timeline

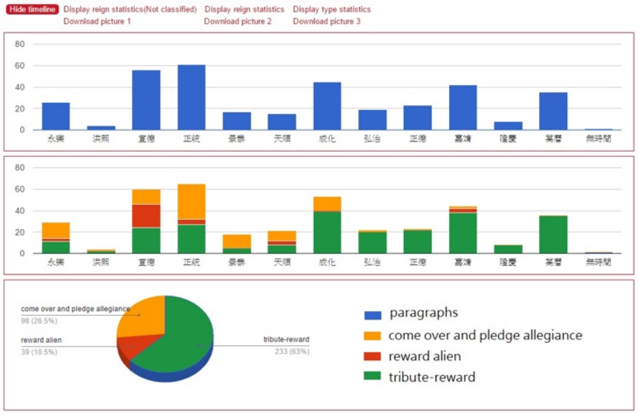
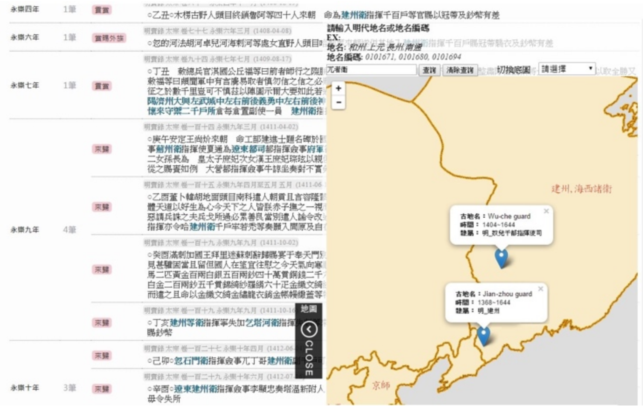


Figure 3. Data visualization options

3. Search snippets: Text snippets of paragraphs related to the searched event types are displayed in window c (Figure 1, (c)). The results are organized in a table with columns (L-R) showing time, number of paragraphs, event type, and related paragraph snippets. All guard mentions in the texts are highlighted and linked to Academia Sinica's CCTS-API Map Service . When a user clicks on a guard link in the text, the guard's location will be shown on a map of Ming China in a pop-up window, see Fig. 4. It shows the locations of Wu-che guard (兀者衛) and Jian-zhou guard (建州衛) in the same map.



永樂四年	1筆	賞賞	○乙丑○木楞古野人頭目絃鎖魯阿等四十人來朝　命為建州衛指揮千百戶等官賜以冠帶及鈔幣有差
永樂六年	1筆	賞賜外族	明實錄 太宗 卷七十七 永樂六年三月 (1408-04-08) ○忽的河法胡河卓兒河海剌河等處女直野人頭目哈
永樂七年	1筆	賞賞	明實錄 太宗 卷九十四 永樂七年七月 (1409-08-17) ○丁丑　敕總兵官淇國公丘福等曰前者師行之際 敕福等曰傾聞軍中有言虜易取者慎勿信之信之必 征之於數千里豈可不慎茲以陣圖示爾大要如此若 陽濟州大興左武城中左右前後義勇中左右前後神 懷來守禦二千戶所倉每倉置副使一員　建州衛指
永樂九年	4筆	來歸	明實錄 太宗 卷一百十四 永樂九年三月 (1411-04-02) ○庚午安定王尚旀來朝　命工部建進士題名碑於國 事薊州衛指揮使夏通為遼東都司都指揮僉事府軍 二女孫長為　皇太子庶妃次女漢王庶妃琮玹以親 從之賜賚如例　大營都指揮僉事牛諒坐奏對不實
		來歸	明實錄 太宗 卷一百十五 永樂九年四月至五月 五月 (1411-06 ○乙酉董卜韓胡地面頭目南科遣人朝貢且言容隆 體天道以好生為心今天下之人皆朕赤子撫之一視 惡請兵誅之夫兵戈所過必累善良當別遣人諭令改 指揮亦令哈建州衛千戶牢若禿等奏願入開原及自
		來歸	明實錄 太宗 卷一百十九 永樂九年九月 (1411-10-02) ○癸酉滿剌加國王拜里迷蘇剌辭歸賜宴于奉天門 見甚驩固當且留但國人在鋻宜往慰之今天氣向寒 馬二匹黃金百兩白銀五百兩鈔四十萬貫銅錢二千 白金二百兩鈔五千貫錦綺紗羅絹六十疋金織文綺 而遣之且命以金織文綺金繡龍衣銷金帳幔繖蓋等
		來歸	明實錄 太宗 卷一百十九 永樂九年九月 (1411-10-16 ○丁亥建州等衛指揮筝失加乞塔河衛指揮 賜鈔幣
永樂十年	3筆	來歸	明實錄 太宗 卷一百二十七 永樂十年四月 (1412-06-0 ○己卯○忽石門衛指揮僉事兀丁哥建州衛
		來歸	明實錄 太宗 卷一百二十九 永樂十年六月 (1412-07- ○辛酉○遼東建州衛指揮僉事李顯忠奏塔溫新附人 毋令失所

請輸入明代地名或地名編碼
EX:
地名: 和州,上元,長州,南益
地名編碼: 0101671, 0101680, 0101694

兀者衛 查詢 清除查詢 切換底圖 請選擇

古地名：Wu-che guard
時間：1404~1644
隸屬：明_奴兒干都指揮使司

古地名：Jian-zhou guard
時間：1368~1644
隸屬：明_建州

Figure 4. Academia Sinica CCTS-API map service

Finally, we conduct a case study targeting Jurchens subordinated garrisons, including Wu-che guards (兀者諸衛) , Jian-zhou guards (建州諸衛), Mao-lian guard (毛憐衛) by using the proposed tool to obtain statistics regarding tribute event types.

We compare our event classification results with Cheng's study (N. Cheng, 2015), which used Ming Shilu as the source to investigate the tribute events during Yongle (永樂), Hongxi (洪熙), and Xuande (宣德) periods. We regard the paragraphs categorized as "tribute-reward", "come over and pledge allegiance", and "reward alien" event types as those potentially illustrating tribute events and manually check them. For Wu-che, Jian-zhou, and Mao-lian guards, 69, 86, and 40 paragraphs are identified as the above three types, respectively. Among these paragraphs, 66, 77, 37 are correct, which are close to the numbers of tribute events in Cheng's study for these three guards (60+, 70+, 30-). This study was done within 16 man-hours. These preliminary results are consistent with Cheng's manual analysis results and show that our tool not only helps historians study Weiso events more efficiently but also keep the quality.

Bibliography

Le, Q. V. and Mikolov, T. (2014). *Distributed Representations of Sentences and Documents.* Beijing, China, pp. 1188–96.

Chinea-Rios, M., Sanchis-Trilles, G., and Casacuberta, F. (2015). *Sentence clustering using continuous vector space representation.* Santiago de Compostela, Spain, pp. 432–40.

Cheng, N. (2015). "A Study of the Tributary System of Jurchens in the Ming Dynasty." *Journal of Chinese Humanities*, 347: 90-109+166-167.

Posters

VIA: Video-dance, computer-assisted composition and mobile technology

Daniella Aguiar
daniella.aguiar@gmail.com
Federal University of Uberlandia, Brazil

Joao Quieroz
queirozj@gmail.com
Federal University of Juiz de Fora, Brazil

Luiz Casteloes
lecasteloes@gmail.com
Federal University of Juiz de Fora, Brazil

VIA is a mobile art project combining video-dance, computational music and architecture. Its main goal is to endow public locations of downtown Rio de Janeiro with video-dance and music, accessed through locative media (smartphones or tablets). The VIA website is the main way of learning about the project, as it provides all the necessary information, including a map that indicates the specific locations where users are able to access the multimedia pieces of the project. Two "Vias", or routes, can be selected on the map, "Via 1" or "Via 2". Each of them is a distinct route on the map linking specific locations according to landscape features (as we shall explain in the next section). Although the project suggests experiencing the work according to these two routes, each viewer can decide how and when they are going to do so. There is not only one precise route to experience the project: some may decide to access the content in a single location and then leave, whereas others might want to visit all the locations creating their own routes from one point to another. Besides the website, the public can also learn about the project through printed maps distributed in Rio de Janeiro's downtown areas where VIA is located. The distributed maps are the same as the website's, except for the fact that there is not a distinction between Vias 1 and 2. It also includes the addresses of the locations in order to guide viewers through the project. The main technology employed to access the project is the QR Code. QR Code is a specific two-dimensional barcode target that can be read by barcode readers and phone cameras. Once viewers decide on a location to start, they must walk to it, where they will find QR Codes on walls, utility poles, litter bins and others. With a QR code reader application, they must frame the QR barcode and wait for the content to be loaded. Summarily, any user equipped with a tablet or smartphone with an Internet connection has free access to experimental multimedia pieces of video-dance and music while moving through specific points of Rio de Janeiro. The multimedia content resulted from the collaboration between João Queiroz (artistic director), Daniella Aguiar (dancer and choreographer), Luiz E. Castelões (music composer), Adriano Mattos Corrêa (architect), Guilherme Landin and Claudia Rangel (video-makers), and Alfredo Suppia (video editor). Each multimedia product was the result of artistic investigation taking place between dance, music, and the architectural richness of Rio de Janeiro urban space, providing pedestrians with an experience that superposes "navigation" through the city and contemporary dance and music. The sounds/music accompanying these videos derive from CAC (Computer-Assisted Composition), CGA (Computer-Generated Assistance), and Sonification – related approaches. They consist of image-to-sound conversions using patches developed in OpenMusic. These conversions employ, as input images, a collection of photographs previously taken from the locations where the dancing took place. The compositional action involved consists mainly of normalizing the xy axis data extracted from the contours of these images within audible ranges. There has been no further compositional manipulation of the input data – such as displacement, editing, looping, etc. Two types of location were chosen to create the "vias". In the first type, the chosen locations, although varying from open spaces, as squares, to narrow streets, share the property of being places with ongoing movement of people. The second type is characterized by building interspaces, independent of other architectural characteristics. The dance movements were developed in strict relation to the chosen spots. In this way, the dance is created in relation both to the buildings and to the passing people. The audio-visual language is based on discontinuous approaches and withdrawals from the "theme" (performer), which irregularly repeats short, alternated and cyclic movements. The video editing is rigorously metric and based on the organization of parts in regularly juxtaposed sequences. This method fixates and highlights invariant properties of the landscape algorithmically translated by the music. One of the project's main features is that viewers can access the pieces at the exact same location where the video and the music were created. As a result, mobile technology users connected to the Internet are able to access the multimedia pieces of the project, experiencing environments in which information and virtual objects "overlap" physical reality.

Getting Medieval: Open Access and Networked Pedagogy

Suzanne Akbari
s.akbari@utoronto.ca
University of Toronto, Canada

Alexandra Bolintineanu

alexandrabolintineanu@gmail.com
University of Toronto, Canada

Getting Medieval: The Many Middle Ages is an undergraduate course, a digitally-inflected introduction to the global Middle Ages. *Getting Medieval* functions as an experiment in networked pedagogy, drawing on the expertise of faculty researchers, curators, and digital scholarship specialists. Students read medieval texts, handle related medieval artifacts from the University's special collections, and curate digital collections and exhibits that bring together texts, artifacts, and digital repositories from around the world. Students also attend biweekly guest lectures by faculty researchers and cultural heritage professionals. Our course is predicated on open access--to a wider Middle Ages than the traditional western European canon reveals; to open-source, open-access platforms for public scholarship and digital collections; to a technical infrastructure open to members of the university community; and to the fragile artifacts of the medieval past from the vaults of local cultural heritage institutions.

We conceived the course as a series of moments in time and space, moments defined by literary texts and related physical objects—when possible, objects from the University of Toronto's Malcove Collection, which students can visit in person. For example, the unit on Geoffrey Chaucer's Canterbury Tales invited students into fourteenth-century England, along the road from London to Canterbury. Students learned about medieval pilgrimage and religious practices around relics. In this context, they encountered Chaucer's pilgrims—and especially the false-relic-peddling Pardoner. And in this context, too, they touched and handled a twelfth-century French reliquary from the University of Toronto's Art Centre. The course focuses on a wide variety of times, places, and literary texts, beyond the traditional western European focus of Medieval Studies: from eighth century Anglo-Saxon England and the Old English *Beowulf* to tenth century medieval Spain and the interweaving of Jewish, Muslim, and Christian theological and poetic traditions in the region of Al-Andalus; from the eleventh-century Japanese imperial court and Murasaki Shikibu's courtly *Tale of Genji* to the twelfth-century Anglo-Norman court and the lays of Marie de France; from the cultural crossroads of the Ethiopian capital of Axum in the thirteenth century to the cultural crossroads of the French capital of Paris in the fourteenth; from the East-West journeys of Asian, Middle-Eastern, and European travelers along the Silk Road, to the journey of Geoffrey Chaucer's pilgrims from London to Canterbury. To do these varied traditions justice, we invited guest speakers to introduce their research. In one class, students listened to an Old English lexicographer talk about linguistic archaeology at the Dictionary of Old English. In another, they learned about a digital collection of medieval Tibetan letters, and they turned the pages of a twentieth-century Tibetan book which maintains medieval book forms. In yet another, students experienced a lecture that turned into a performance of medieval drama. The lectures provide access not only to the usual undergraduate curriculum, but to a range of faculty-led research projects that showcase the range of Medieval Studies at Toronto.

The digital inflection of the course is twofold. First, students examine how the times, places, texts, and things we study are represented in digital projects and repositories, from the Dictionary of Old English to the International Dunhuang Project. Second, students are invited to create public-facing scholarship themselves: After they visit the secure storage spaces of the Malcove Collection and receive the curator's guidance on the care, handling, and cultural context of artifacts, students handle medieval artifacts; study them in the context of our local collection and of digital repositories around the world; and produce a public, curated digital collection and exhibit, linking the artifact to the course's thematics, in the open-source content management platform Omeka. In so doing, students practice close reading and slow looking, examining an object in cultural and literary contexts. They also learn about digital collections—from the digitization of materials to data curation and digital preservation—and about scholarly communication possibilities in the digital age.

We designed our digital assignments with access in mind. The students' initial contact with the secure vaults of the Malcove Collection offered hands-on, immediate, physical access to the artifacts—and included a discussion on the kinds of information such access provides, information unavailable in a digital medium. The software platform Omeka itself is free, open-source, with significant traction in DH pedagogy. Our hosted Omeka instances allow students to complete the work even if they only have access to public university computers. Our tutorials guide students through the work in interactive workshops conducted in university computer laboratories, ensuring that students without a laptop, or without prior technical experience, are able to complete their work in a supportive environment, with help close at hand. But this level of access requires significant logistical work and infrastructure. Students have access to faculty research projects thanks to the Centre for Medieval Studies' network of scholars. Students are able to use university laboratories because we make arrangements in collaboration with digital scholarship librarians to provide access to this portion of UofToronto's computing infrastructure. Students are able to experience relevant artifacts from the Malcove collection so closely because of a course-long collaboration between its Collections Manager, Heather Pigat, and the course instructors.

Thus *Getting Medieval* functions as an experiment in networked pedagogy. Biweekly guest lectures, robustly integrated into the curriculum, invite a network of faculty researchers and cultural heritage professionals into the undergraduate classroom; and the open access requirements of our digital projects draw on the institution's physical and human infrastructure—from special collections of medieval objects to computing facilities; from curators to medievalists to digital scholarship specialists. As we invite our students on a time-travelling journey across the global Middle Ages, we rely on a network of guides to open points of access along the way.

Coffee Zone: Del cafetal al futuro / From the coffee fields to the future

Mark Anderson
mark-anderson@uiowa.edu
University of Iowa, United States of America

Hannah Scates Kettler
hannahskettler@gmail.com
University of Iowa, United States of America

In the last decade, the number of coffee farms in Puerto Rico has shrunk from around 11,000 to nearly 4,000. Economic and climate conditions, as well as the increased migration of young people to the more cosmopolitan coastal areas in pursuit of education, have resulted in significant changes to the island's coffee-growing regions. These changes, especially with regard to linguistic characteristics of the coffee growers themselves, have been the subject of a multi-year study by University of Iowa Linguistics Professor Julia Oliver Rajan.

The Digital Scholarship & Publishing Studio at the University of Iowa collaborates with faculty and students on the digital design, implementation, and circulation of their research. The Studio embraces scholarly creativity and encourages interdisciplinary research and multiplatform circulation. In this manner, the Studio helps scholars tailor the presentation and application of their research to a variety of audiences.

Together with Professor Oliver Rajan, the Studio has created a unique bilingual (in Spanish and English) digital archive of oral history videos – Coffee Zone: Del cafetal al futuro / From the Coffee Fields to the Future documents a vanishing dialect of Spanish spoken in the mountainous coffee growing regions of Puerto Rico. Currently consisting of over 600 short video clips in 16 topical categories, the archive demonstrates how rapid migration to the cities has altered the power structures of coffee plantations, the role of women and young people, and the new landscape of coffee production. The site can serve as a template for other researchers who are documenting similarly endangered languages or dialects in other parts of the world.

We propose to present a poster at DH2017 reporting the progress and challenges of this digital humanities project, how it acts as a resource for scholars and students in a wide variety of disciplines (ecology, horticulture, psychology, and obviously linguistics, just to name a few), and the upcoming features we are working to implement. We will outline the upcoming features and how this specific project, though currently focused on linguistic shifts in the coffee zone, applies to much larger, global concerns of globalized economies, climate change and the loss of language as it relates to cultural identity.

Although the videos are now freely available worldwide, accessibility is limited because they are neither transcribed nor translated from their original Spanish. Closed captions in the videos would enhance their usefulness to a wider audience, but a major challenge lies in the very nature of the oral histories – the unique dialect of Spanish spoken by the interviewees. To overcome this, we look to expand the project by collaborating with the University of Puerto Rico, where they would contribute to it by assisting with the tasks of transcription, translation, and caption encoding as well as test the site in their courses. Another feature we look to add is an interactive map showing the geographic areas affected by these changes, which would function as an additional access point to the videos themselves.

Distant Seeing TV

Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

We present work on *Distant Seeing TV*, which applies computational techniques to the study of television series. Our initial set of interest consists of five American sitcoms spanning the 1950's through the 1970's. In order to focus the academic questions that we are interested in exposing, all of the selected shows feature leading women characters and have previously been the topic of academic studies. The poster will outline the kinds of academic questions we hope to answer with our study, the computational methods currently available for answering these questions, our novel extensions of these methods, and some initial results.

The project *Distant Seeing TV* brings the approach of Moretti's "Distant Reading" and Manovich's cultural analytics to the analysis of a large corpus of moving images (Moretti 2013, Manovich 2001). Given that long-running television series broadcast hundreds of episodes and the major networks run dozens of series each season, previous studies, including (Baughman 1993), (Dow 1996), (Spangler 2003), (Morely 2003), have largely had to rely on a close analysis of a small subset of series, episodes, and even scenes. Our project seeks to augment these approaches with computationally-driven analyses that can curate and aggregate the contents of tens of thousands of hours of television programming. To do this, we extract features such as the placement and identities of faces in the shot, as shown in Figure 1, and the time codes for laughter, music, and scene changes as well as the identity of scene locations, as seen in Figure 2. There has been relatively little work done on an aggregate analysis of a corpora of moving images. Noticeable exceptions include *Cinemetrics* (Tsivian and Civjans 2011), which focuses on average shot length in cinema, and the *Arclight Guidebook to Media History and the Digital Humanities* (Acland and Hoyt 2016). Our work extends these to a much wider set of metrics and centers on issues specific to television in contrast to those of film.

Our work attempts to address several types of questions of interest within television studies. These include: Are women characters seen entering a scene more frequently than men? Which characters tend to walk in front of other characters? Does the typical sequence of locations change throughout the run of a show or across different shows? How can we best characterize the narrative arc of an episode? These questions address issues of auteurship, gender, race and narrative in TV.

The initial corpus we are working with includes a diverse set of series with women lead characters that span the 1950s through 1970s. They include a mix of episodes filmed in black and white, in color, with a multiple-camera set-up, with a single camera set up, and from all three major networks:

- I Love Lucy, 1951-1957 (b/w; multiple-camera; CBS)
- The Donna Reed Show, 1958-1966 (b/w; single camera; ABC)
- Bewitched, 1964-1972 (b/w to color; single camera; ABC)
- I Dream of Jeannie, 1965-1970 (b/w to color; single camera; NBC)
- The Mary Tyler Moore Show, 1970-1971 (color; multiple-camera; CBS)

Focusing on a small but diverse set helps in the early stages on this project as we adaptively learn what tools and techniques are most interesting and explore how these methods intersect with current scholarship.

There has been a rapid rate of advancement in modern computer vision techniques over the past decade, which has been particularly driven by the use of deep convolutional neural networks. Given their novelty, there are limited computer vision tools that can be used directly out of the box on a new problem domain. Much of our work on the project has been focused on building a tool set specifically engineered for analyzing television, with a particular focus on parsing black and white images. We will release these tools as an open source package, with wrappers written in python, once it becomes stable enough to be used by other researchers. Techniques that have or will implement include facial detection and recognition (Sun, et al. 2015), scene break detection (Kar 2015; Kumar, Gupta, and Venkatesh 2015; Pulvar 2015), scene classification (Cheng 2015), dialogue disambiguation (Cervone 2015), speech detection (Sanders, Taubman, and Lee 2015), audience and laugh track detection (Cosentino 2015; Joshi 2016), music segmentation, camera angle detection, and camera selection detection (Nadimi and Bhanu 2004). In all of these cases, training specifically on historical television data has yielded significantly better results that models trained on generic corpora.

The study of television has often been downplayed in favor of textual sources and feature-length film. As described in (Fiske 1978): "Television suffered categorical disadvantage in repute ... its characteristic was oral not literate, whereas 'dominant culture' ... was militantly committed to print-literacy and the values associated with that." From the 1950's onward, however, television has arguably served as the dominant source of mass entertainment in the United States. By 1959, over 83% of households in the US owned their own television set (Baughman 1993). This is not to say that there has been no substantive research in television studies. In fact, there is a large set of prominent examples, including (Baughman 1993), (Dow 1996), (Spangler 2003), (Morely 2003), and many more. Given that long-running television series broadcast hundreds of episodes and the major networks run dozens of series each season, these studies have largely had to rely on a close analysis of a small subset of series, episodes, and even scenes. Our project seeks to augment these approaches with computationally-driven analyses that can curate and aggregate the contents of tens of thousands of hours of television programming.

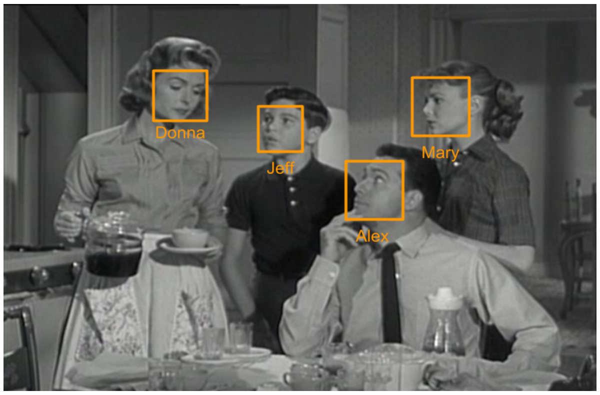


Figure 1. An example of face detection and disambiguation from a still image taken from The Donna Reed Show. Tracking these over the course of a scene and episode reveal characteristics of how characters interact and describe the narrative flow of the series.

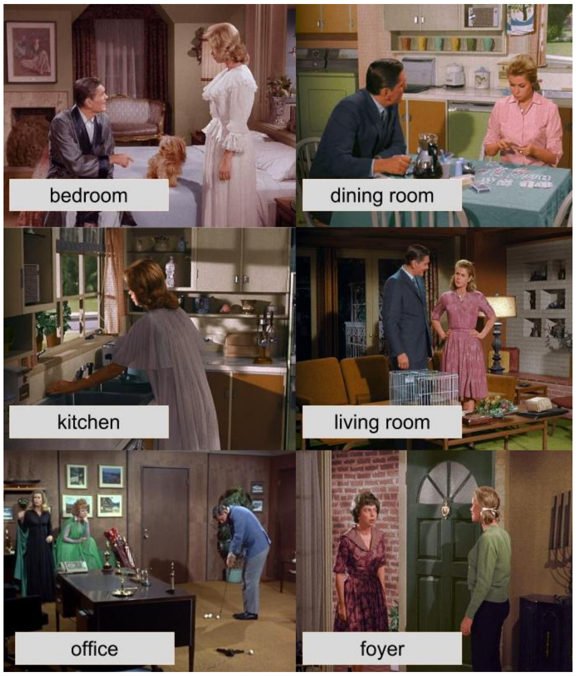


Figure 2. An example of scene classification from Bewitched. Each still image from the episode is tagged with the description of the sets on the sound stage. Following the progress of these over the course of the episode can serve as a way to compactly describe and aggregate the narrative arc of an episode. Comparing across episodes, seasons and shows reveals similarities and differences across the various series of interest.

Bibliography

Acland, C. R. and Hoyt, E. (2016) Editors. *The Arclight Guidebook to Media History and the Digital Humanities.* REFRAME Books.

Baraldi, L., Grana, C., and Cucchiara, R. (2015). "Measuring Scene Detection Performance." *Iberian Conference on Pattern Recognition and Image Analysis*. Springer International Publishing.

Baughman, J. L. (1993) "Television Comes To America, 1947-1957'." *Illinois History* 46.3.

Baughman, J. L. (2006) *The Republic of Mass Culture: Journalism, Filmmaking, and Broadcasting in America Since 1941*. JHU Press, 2006.

Cervone, A., et al. (2015) "Towards automatic detection of reported speech in dialogue using prosodic cues." *Sixteenth Annual Conference of the International Speech Communication Association*.

Cheng, G., et al. (2015) "Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images." *IET Computer Vision* 9.5: 639-647.

Cosentino, S., et al. (2015) "Automatic discrimination of laughter using distributed sEMG." *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE.

Dow, B. J. (1996) *Prime-Time Feminism: Television, Media Culture, and the Women's Movement Since 1970*. University of Pennsylvania Press.

Fiske, J. (1978). *Reading Television*. Routledge.

Joshi, A., et al. (2016) "Harnessing Sequence Labeling for Sarcasm Detection in Dialogue from TV Series 'Friends'." *CoNLL 2016*: 146.

Kar, T., and Kanungo, P. (2015). "A texture based method for scene change detection." *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*. IEEE.

Kumar, R., Gupta, S., and Venkatesh, K. S. (2015) "Cut scene change detection using spatio temporal video frame." *2015 Third International Conference on Image Information Processing (ICIIP)*. IEEE.

Pulver, A., Chang, M-C., and Lyu, S. (2015) "Shot Segmentation and Grouping for PTZ Camera Videos." *10th Annual Symposium on Information Assurance (ASIA'15)*.

Manovich, L. (2001). *The Language of New Media*. MIT press.

Moore, B., Bensman, M. R., and Van Dyke, J. (2006). *Prime-Time Television: A Concise History*. Greenwood Publishing Group.

Morley, D. (2003). *Television, Audiences and Cultural Studies*. Routledge,

Moretti, F. (2013). *Distant reading*. Verso Books.

Nadimi, S., and Bhanu, B. (2004) "Physical models for moving shadow and object detection in video." *IEEE transactions on pattern analysis and machine intelligence* 26.8 (2004): 1079-1087.

Sanders, J., Taubman, G., and Lee, J. J. (2015). "Background audio identification for speech disambiguation." U.S. Patent No. 9,123,338. 1 Sep. 2015.

Silverstone, R. (1994) *Television and everyday life*. Routledge, 1994.

Spangler, L. C (2003). *Television Women from Lucy to Friends: Fifty Years of Sitcoms and Feminism*. Greenwood Publishing Group, 2003.

Sun, Y., et al. (2015) "Deepid3: Face recognition with very deep neural networks." *arXiv preprint arXiv:1502.00873* (2015).

Tsivian, Y., and Civjans, G. (2011) "Cinemetrics: Movie Measurement and Study Tool Database." (2011).

Xu, P., et al (2001). "Algorithms And System For Segmentation And Structure Analysis In Soccer Video." *ICME*. Vol. 1. 2001.

1947 Partition on the Margins: The Untold Testimonies of Sikh, Bahawalpur and Marwari Communities

Shaifali Arora
arora.shaifali16@gmail.com
School of Humanities and Social Sciences
Indian Institute of Technology Indore, India

Nirmala Menon
nmenon@iiti.ac.in
School of Humanities and Social Sciences,
Indian Institute of Technology Indore, India

This poster offers a concise working plan of a digital project titled *1947 Partition on the Margins: The Untold Testimonies of Sikh, Bahawalpur and Marwari Communities*. The project aims to collect undocumented narratives of many

'Punjabi' and 'Rajasthani' communities who migrated from villages in current Eastern Pakistan to villages in the current Western India during the 1947 partition. The primary objective of the project is to create an open access visual and narrative based digital archive of partition and post-partition testimonies developed through audio-video transcripts and photographs. The audio-visual transcripts will be collected through personal interviews of the partition witnesses, their families and subsequent generations. These interviews will take place on-site in their villages. This population mostly consists of Sikh, Bahawalpur and Marwari communities who travelled comparatively short distances during partition, settling in the nearby villages close to the border in Indian states of Rajasthan and Punjab. The geographical distance travelled by these communities towards Indian villages was only about 40 to 50 KM. However, the spatial definitions changed and any hope of revisiting the ancestral homeland was shattered with the newly constructed international border.

These are a few minority communities residing in small villages whose partition experiences are not traced in public memory or documented in any literature, primarily because their journey of migration was relatively hushed and non-violent. However, the geo-political movement from one place to another brought larger social, cultural, political and economic challenges in the post-independence India for families in these many communities. They migrated from one small village to another leaving behind whatever geographical, economical and cultural assets that belonged to them. Even after almost 70 years of partition in 1947, families in these communities are struggling to own lands of their own. The generation that witnessed the partition of 1947 is disappearing and will not be there after next 4 to 5 years. Therefore, it becomes an obligatory task to document the narratives of a traumatic historical event that completely changed the flow of history.

The subsequent generations of these families who migrated to India continue to face economic and cultural challenges. The nostalgia of not being able to visit their ancestral homeland disturbs the process of connecting with their current homeland. Therefore, this project aims to contextualise these first-hand audio-video interviews and photographs into narratives that will provide an influential palimpsest of history and the contemporary cultural and political state.

There are several ongoing digital projects that contextualise historical events into narratives such as *1947 Partition Archive*, *1984 Living History*, *Kashmir Oral History*, *Indian Memory Project* and *Holocaust Memorial Museum*. The memory projects on the 1947 partition such as *1947 Partition Archive* mostly record stories that are often (rightly so) narratives of violence and bloodshed. However, there is gap in existing literature and database that do not document less violent partition experiences of a large population that mostly live at the western Indian border. The aim of the current project is to document those unheard communities which escaped immediate violence but continue to suffer the residual pains of geo-cultural and economical traumas in post-partition India. The poster also offers a framework of methodology to be followed for developing the project. The methodology involves collection of data by visiting the villages, audio-video interviews with families, collecting photographs and letters, transcription and description of interviews and contextualising photographs and letters in narratives. The development of final testimonies achieved through these working stages will be developed into an open access website, hosted through KSHIP (Knowledge Sharing In Publication) which is a multilingual academic publishing project at IIT (Indian Institute of Technology) Indore, India.

Athar: Numérisation, indexation et mise en ligne enrichie des bulletins du Comité de conservation des monuments de l'art arabe (1882–1953)

Hélène Bégnis
helene.begnis@persee.fr
Université Lyon, ENS de Lyon, CNRS, France

Antonio Mendes da Silva
antonio.mendes@inha.fr
INVISU, Institut National d'Histoire de l'Art, France

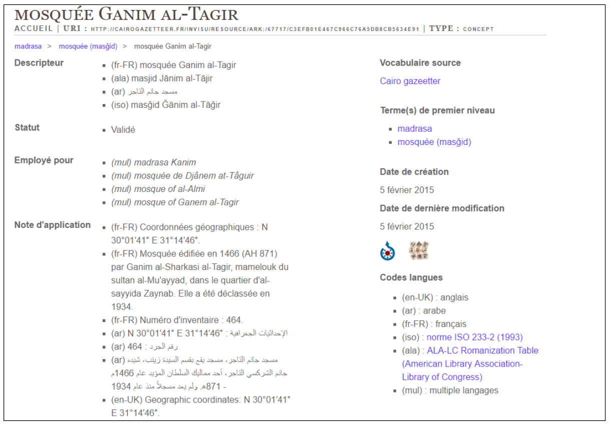
Le projet Athar (de l'arabe traces ou empreintes, et par extension monuments ou antiquités) met à disposition, en open access, l'intégralité des travaux du *Comité de conservation des monuments de l'art arabe* (1882-1953), instance créée au Caire pour inventorier, décrire et restaurer les monuments islamiques puis coptes d'Égypte. Sur le modèle des *rapports de la Commission des monuments historiques en France*, cette publication annuelle, rédigée en français, aujourd'hui éteinte et libre de droits, se compose de 41 volumes publiés au Caire entre 1882 et 1953, totalisant 8 000 pages et près de 800 planches. Ce corpus rassemble parfois l'unique documentation encore subsistante sur des monuments disparus, très dégradés ou radicalement transformés. Par ses illustrations, il se trouve en relation avec les riches fonds de la photographie commerciale présente au Caire à partir de 1860, diffusée de façon croissante sur Gallica et Europeana.

Les partenaires

Ce projet est le résultat d'une collaboration fructueuse entre trois structures du monde scientifique et universitaire français : le laboratoire InVisu, l'Institut français d'archéologie orientale et Persée. La complémentarité des expertises de chacun de ces partenaires a été essentielle pour mener à bien ce projet. Le laboratoire InVisu, grâce à son savoir-faire documentaire, a pris en charge la structuration fine du corpus et l'indexation des toponymes à l'aide du *Cairo Gazetteer*. Ce référentiel multilingue (français, arabe, anglais) permet d'identifier, de décrire et de localiser finement les 600 édifices classés du Caire au titre de monuments historiques en proposant les variantes orthographiques de ces toponymes. L'IFAO a permis la complétion de la collection en prêtant les volumes manquants et a apporté ses compétences sur la toponymie et la cartographie égyptiennes. Persée, grâce à son expertise de plus de 10 ans en matière de numérisation, d'océrisation et de diffusion du patrimoine scientifique, a mis au service du projet sa plateforme technique (JGAlith : application intégrée et complète, développée par Persée, qui permet à l'ensemble de ses partenaires impliqués dans le processus de traitement des collections de communiquer et de partager les informations apportées par chacun) et ses ressources humaines assurant ainsi la gestion de l'ensemble du cycle des documents (dématérialisation, documentation, diffusion web, sauvegarde et archivage pérenne).

Technologies et formats, usage du Linked Open Data

Les technologies et formats utilisés reposent sur les normes et les standards en vigueur dans le contexte de l'open access et de l'interopérabilité. Les métadonnées générées par la structuration du corpus sont encodées au format XML selon les schémas Dublin Core, MarcXML et MODS. Le texte intégral issu de l'OCR non corrigé est disponible selon le schéma TEI et l'ensemble des données est décrit et organisé dans un container XML au format METS. Le contenu du *Cairo Gazetteer*, déployé grâce à Ginco (outil de gestion de référentiels terminologiques), est mis à disposition de la communauté des humanités numériques par le biais du format SKOS. Il offre une solution innovante pour résoudre les difficultés liées à la translittération des toponymes arabes.



MOSQUÉE GANIM AL-TAGIR
ACCUEIL | URI : HTTP://CAIROGAZETTEER.FR/INVISU/RESOURCE/ARK:/67717/C3EFB01E467C966C76A5DB8CB5634E91 | TYPE : CONCEPT

madrasa > mosquée (masǧid) > mosquée Ganim al-Tagir

Descripteur
- (fr-FR) mosquée Ganim al-Tagir
- (ala) masjid Jānim al-Tājir
- (ar) مسجد جانم التاجر
- (iso) masǧid Ǧānim al-Tāǧir

Statut
- Validé

Employé pour
- *(mul) madrasa Kanim*
- *(mul) mosquée de Djânem at-Tâguir*
- *(mul) mosque of al-Almi*
- *(mul) mosque of Ganem al-Tagir*

Note d'application
- (fr-FR) Coordonnées géographiques : N 30°01'41" E 31°14'46".
- (fr-FR) Mosquée édifiée en 1466 (AH 871) par Ganim al-Sharkasi al-Tagir, mamelouk du sultan al-Mu'ayyad, dans le quartier d'al-sayyida Zaynab. Elle a été déclassée en 1934.
- (fr-FR) Numéro d'inventaire : 464.
- (ar) N 30°01'41" E 31°14'46" : الإحداثيات الجغرافية
- (ar) 464 : رقم الجرد
- (ar) مسجد جانم التاجر، مسجد يقع بقسم السيدة زينب، شيده جانم الشركسي التاجر، أحد مماليك السلطان المؤيد عام 1466م - 871هـ. ولم يعد مسجلاً منذ عام 1934
- (en-UK) Geographic coordinates: N 30°01'41" E 31°14'46".

Vocabulaire source

Cairo gazeetter

Terme(s) de premier niveau
- madrasa
- mosquée (masǧid)

Date de création

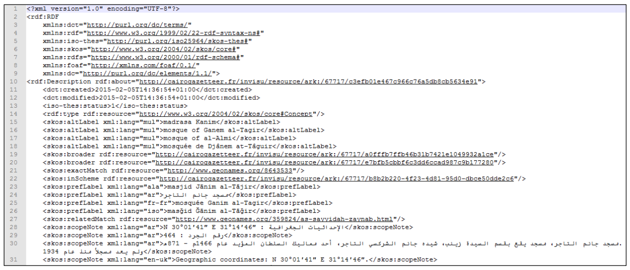
5 février 2015

Date de dernière modification

5 février 2015

Codes langues
- (en-UK) : anglais
- (ar) : arabe
- (fr-FR) : français
- (iso) : norme ISO 233-2 (1993)
- (ala) : ALA-LC Romanization Table (American Library Association-Library of Congress)
- (mul) : multiple langages

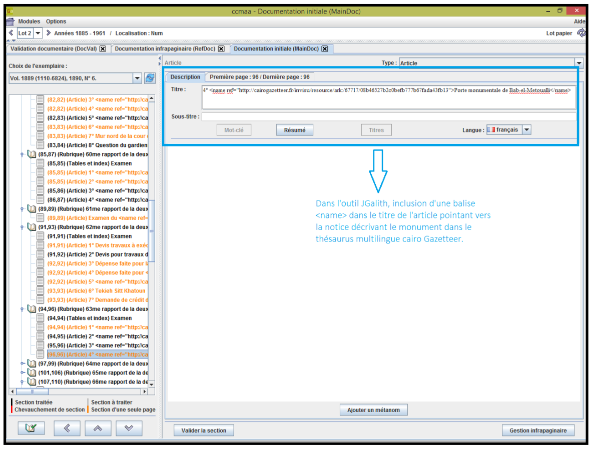
Illustration 1. Une fiche du Cairo Gazetteer



```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:dct="http://purl.org/dc/terms/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:iso-thes="http://purl.org/iso25964/skos-thes#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://cairogazetteer.fr/invisu/resource/ark:/67717/c3efb01e467c966c76a5db8cb5634e91">
    <dct:created>2015-02-05T14:36:54+01:00</dct:created>
    <dct:modified>2015-02-05T14:36:54+01:00</dct:modified>
    <iso-thes:status>1</iso-thes:status>
    <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
    <skos:altLabel xml:lang="mul">madrasa Kanim</skos:altLabel>
    <skos:altLabel xml:lang="mul">mosque of Ganem al-Tagir</skos:altLabel>
    <skos:altLabel xml:lang="mul">mosque of al-Almi</skos:altLabel>
    <skos:altLabel xml:lang="mul">mosquée de Djânem at-Tâguir</skos:altLabel>
    <skos:broader rdf:resource="http://cairogazetteer.fr/invisu/resource/ark:/67717/a0fffb7ffb46b31b7421e1049932a1ce"/>
    <skos:broader rdf:resource="http://cairogazetteer.fr/invisu/resource/ark:/67717/e7bfb5cbbf6c3dd6ccad987c9b177280"/>
    <skos:exactMatch rdf:resource="http://www.geonames.org/8643533"/>
    <skos:inScheme rdf:resource="http://cairogazetteer.fr/invisu/resource/ark:/67717/b8b2b220-4f23-4d81-95d0-dbce50dde2c6"/>
    <skos:prefLabel xml:lang="ala">masjid Jānim al-Tājir</skos:prefLabel>
    <skos:prefLabel xml:lang="ar">مسجد جانم التاجر</skos:prefLabel>
    <skos:prefLabel xml:lang="fr-fr">mosquée Ganim al-Tagir</skos:prefLabel>
    <skos:prefLabel xml:lang="iso">masǧid Ǧānim al-Tāǧir</skos:prefLabel>
    <skos:relatedMatch rdf:resource="http://www.geonames.org/359824/as-sayyidah-zaynab.html"/>
    <skos:scopeNote xml:lang="ar">N 30°01'41" E 31°14'46" : الإحداثيات الجغرافية</skos:scopeNote>
    <skos:scopeNote xml:lang="ar">464 : رقم الجرد</skos:scopeNote>
    <skos:scopeNote xml:lang="ar">مسجد جانم التاجر، مسجد يقع بقسم السيدة زينب، شيده جانم الشركسي التاجر، أحد مماليك السلطان المؤيد عام 1466م - 871هـ.
    ولم يعد مسجلاً منذ عام 1934</skos:scopeNote>
    <skos:scopeNote xml:lang="en-uk">Geographic coordinates: N 30°01'41" E 31°14'46".</skos:scopeNote>
```

Illustration 2. Les données du Cairo Gazetteer encodées en RDF

Des liens vers deux référentiels disponibles au niveau international ont été insérés dans les données de structuration du corpus : IdRef pour les auteurs et le *Cairo Gazetteer* pour les édifices. Ces liens permettent de déduire des alignements vers des ressources extérieures complémentaires telles que data.bnf.fr, DBpedia, les Subject Headings de la Library of Congress et GeoNames. Le développement et l'implémentation par l'équipe informatique de Persée d'un module spécifique dans la plateforme de JGALITH a été nécessaire pour réaliser cette étape.



ccmaa - Documentation initiale (MainDoc)

Lot 2 › Années 1885 - 1961 / Localisation : Num

Choix de l'exemplaire : Vol. 1889 (1110-6824), 1890, N° 6.

Type : Article

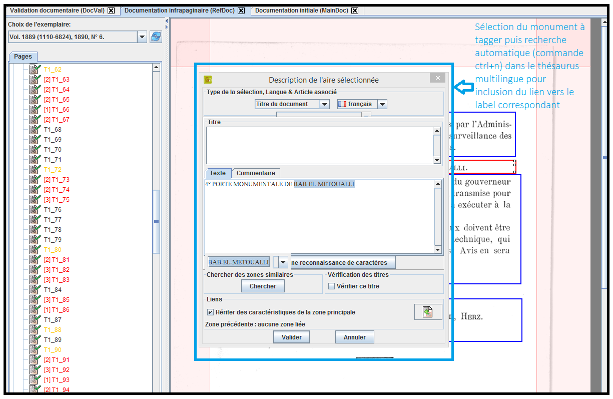
Titre : 4° <name ref="http://cairogazetteer.fr/invisu/resource/ark:/67717/08b46527b2c0befb777b67fada43fb13">Porte monumentale de Bab-el-Metoualli</name>

Langue : français

Dans l'outil JGalith, inclusion d'une balise <name> dans le titre de l'article pointant vers la notice décrivant le monument dans le thésaurus multilingue cairo Gazetteer.

Modules
Options
Aide
Lot papier
Validation documentaire (DocVal)
Documentation infrapaginaire (RefDoc)
Documentation initiale (MainDoc)
Description
Première page : 96 / Dernière page : 96
Sous-titre :
Mot-clé
Résumé
Titres
Ajouter un métanom
Valider la section
Gestion infrapaginaire

Illustration 3. Dans JGalith, inclusion d'une balise <name> dans le titre de l'article pointant vers l'entité nommée correspondante dans le Cairo Gazetteer



Description de l'aire sélectionnée

Type de la sélection, Langue & Article associé : Titre du document ; français

Texte : 4° PORTE MONUMENTALE DE BAB-EL-METOUALLI .

BAB-EL-METOUALLI

☑ Hériter des caractéristiques de la zone principale

Zone précédente : aucune zone liée

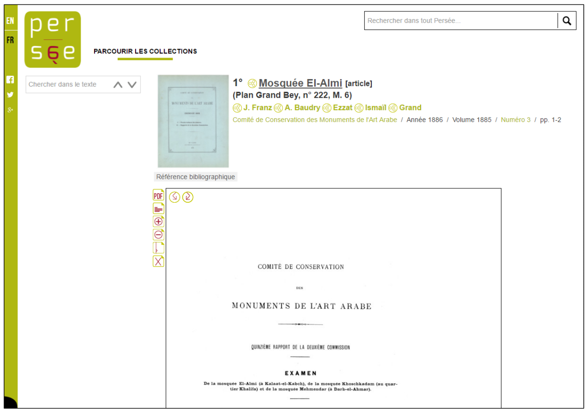
Sélection du monument à tagger puis recherche automatique (commande ctrl+n) dans le thésaurus multilingue pour inclusion du lien vers le label correspondant

Validation documentaire (DocVal)
Documentation infrapaginaire (RefDoc)
Documentation initiale (MainDoc)
Choix de l'exemplaire :
Vol. 1889 (1110-6824), 1890, N° 6.
Pages
Titre
Texte
Commentaire
ne reconnaissance de caractères
Chercher des zones similaires
Chercher
Vérification des titres
☐ Vérifier ce titre
Liens
Valider
Annuler

Illustration 4. Dans JGalith, sélection du monument à lier vers le label correspondant dans le Cairo Gazetteer

Diffusion du corpus

La diffusion de ce corpus est actuellement réalisée sur le portail Persée.


EN
FR
per
sée
PARCOURIR LES COLLECTIONS
Rechercher dans tout Persée...
Chercher dans le texte
Référence bibliographique
PDF

1° Mosquée El-Almi [article]
(Plan Grand Bey, n° 222, M. 6)
J. Franz A. Baudry Ezzat Ismaïl Grand
Comité de Conservation des Monuments de l'Art Arabe / Année 1886 / Volume 1885 / Numéro 3 / pp. 1-2

COMITÉ DE CONSERVATION
DES
MONUMENTS DE L'ART ARABE

QUINZIÈME RAPPORT DE LA DEUXIÈME COMMISSION

EXAMEN
De la mosquée El-Almi (à Kalaat-el-Kabch), de la mosquée Khoschkadam (au quartier Khalifa) et de la mosquée Mehmendar (à Darb-el-Ahmar).

Illustration 5. Affichage sur le portail www.persee.fr

A l'horizon 2017-2019 un site corpus sera développé par Persée faisant ainsi évoluer son offre de services. L'ensemble des données produites et agrégées dans le cadre du projet Athar sera intégré dans un écosystème plus large pour offrir une information plus riche, plus complète et centralisée pour l'utilisateur du site dédié. Des pages composites construites à partir des données produites par la chaîne Persée et celles récupérées grâce à l'alignement avec les référentiels extérieurs sélectionnés seront aussi développées. Elles viendront compléter l'éventail des fonctionnalités de navigation et d'exploration (index auteur, table des matières, références croisées, table des illustrations) déjà mis en œuvre dans le portail Persée. Parallèlement à la diffusion web, une exposition des données par le biais d'un entrepôt OAI et d'un SPARQL endpoint sera également déployée.

Bibliographie

Perrin, E. (2015). "Cairo Gazetteer : a thesaurus to identify the monuments of Cairo". *6e journées du réseau Medici. Multilinguisme : frein ou catalyseur de la diffusion scientifique en Europe et en Méditerranée ?* Marseille: Réseau médici. Available at: https://hal.archives-ouvertes.fr/hal-01213678 [Consulté le mars 31, 2017].

Perrin, E. (2014). "Outils, méthodes, corpus : la modélisation des données en SHS. In Journée d'étude web de données et sciences humaines et sociales". Paris. Available at: https://www.inha.fr/fr/agenda/parcourir-par-annee/en-2014/novembre-2014/outils-methodes-corpus.html [Consulté le mars 31, 2017].

Perrin, E. & Mounier, P. (2015). "Structurer, relier et diffuser des données avec les technologies du web sémantique". *Lundis numériques de l'INHA*. Paris. Available at: https://www.inha.fr/fr/agenda/parcourir-par-annee/en-2015/mai-2015/structurer-relier-et-diffuser-des-donnees-avec-les-technologies-du-web-semantique.html [Consulté le mars 31, 2017].

Mapping Borges in the Argentine Publishing Industry (1930–1951)

Nora C. Benedict
ncb3ka@virginia.edu
University of Virginia, United States of America

"A book is not an isolated being: it is a relationship, an axis of innumerable relationships." Jorge Luis Borges's words resonate not only with much of my work on book history, but also with my ongoing digital mapping project of the publishing industry in twentieth-century Argentina. The central aim of my current research is a materially informed look into Borges's process of production in the early part of his literary career (1930-1951): who were his primary publishers? What types of editorial jobs did he hold? Which works were independently printed and which did others fund? What can we tell about his development of aesthetics based on analysis of typography and paper choices? By asking these types of questions about Borges's preferences toward the publication of his own works, we also find that he had a very specific taste for the look and feel of *others'* books and, in one of his early works from this period, makes a snide, and extremely revealing, remark about second-hand bookstores as nothing more than "turgid purgatories." In light of the fact that the places in which a book is published, printed, and even purchased are all key characteristics of that specific volume, I am creating a project with Leaflet and Jekyll that allows me to think deeply about the spatial (and temporal) evolution of the process of publication (and, in turn, the circulation of published works) in Jorge Luis Borges's Argentina (1930-1951).

My project has three distinct layers: mapping the industry, grids of contact, and textual materiality. The first layer uses data from books' colophons and publishers' catalogues to pinpoint the locations of Borges's printers, publishing houses, booksellers, and even places of employment throughout the early part of the twentieth century. The second layer draws on biographical information from the individuals involved in the Argentine book industry to highlight their connections and working relationships. The third and final layer consists of a (visual) catalogue of certain physical aspects of the books in question including their covers, any illustrations, samples of typefaces that certain printers utilized, and any colophons or printers' marks. This data also will include a more general descriptive bibliography of all of Borges's works published between 1930 and 1951, as well as works that he edited, prefaced, or translated during this time. In light of the fact that there are virtually no extant publishers' archives from this historical moment in Argentina and, furthermore, that much of the ephemeral print material related to Borges is in private hands, my aim is to provide researchers and any interested parties with an easily accessible reservoir of data for future projects. Moreover, the specific data collection and interfacing used for this project allow for a deeper understanding of Borges's idea of a book as "an axis of innumerable relationships" and, thus, provide insight into the accessibility of materials, scarcity of resources, aesthetic features, and canon formation in Argentina. In a way, I see this project as something that will not only further material studies (related to the book and publishing history) and digital projects in Southern Cone Literature, but also something that will serve a much wider audience.

Embedding Digital Humanities in a Classics Master Programme

Aurélien Berra
aurelien.berra@gmail.com
Université Paris-Ouest Nanterre, France

Introduction

Pedagogy in the Digital Humanities is now leaving its "bracketed" state – a term used by HIRSCH 2012 to emphasise the fact that this dimension was not given the consideration its practical importance deserves. As programmes and courses are created on a larger scale and increasingly drive institutional strategies, also in Europe (see Sahle, 2013 and the DARIAH Digital Humanities Course Registry), it becomes essential to make comparisons and shared reflections possible.

Since 2014 all students of Greek and Latin languages and literatures at the Université Paris-Ouest Nanterre (France) have been enrolling in a Master programme entitled "Humanités classiques et humani-tés numériques." Each semester features a fully fledged course of Digital Humanities: it is therefore an experiment in embedding Digital Humanities into an existing discipline, or rather into the array of disciplines which constitute the field of Classical studies around its philological backbone.

The aim of this poster is to share the approach I take in designing and teaching these courses, and to reflect on what this experience suggests about digital educational models, in Classics and beyond.

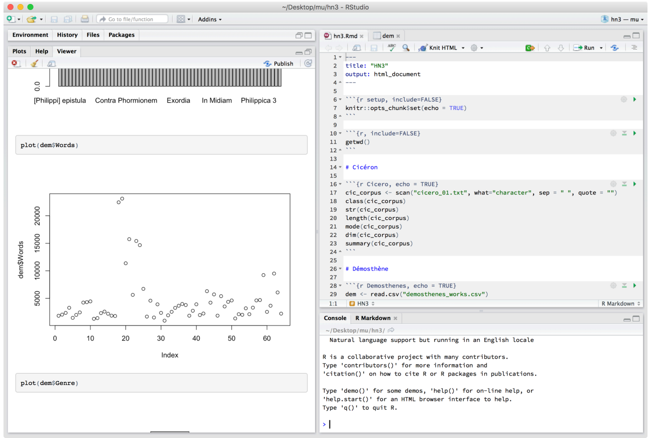
The poster will have three components, devoted to situating, describing and comparing the courses.

Context and History

I will set out the conditions in which the curriculum was reformed (which involves both national and local contexts), the specific problems encountered (as the heterogeneous levels and motivations of the students, the relationships with the other courses, the available technical options, or the recent introduction of podcasting and distance learning), as well as the rationale and methods which shape the courses, including its main sources of inspiration in the Digital Humanities community, whether online syllabi or publications like Jockers (2014) and Rockwell and Sinclair (2016).

Overview of the Courses

The courses alternately take the form of more traditional classes and collaborative or personal pro-jects. Across the two years, their contents include theoretical and historical insights, while concentra-ting on hands-on experience: digital literacy elements are gradually integrated as students go from traditional scholarly editing recreated in Markdown and HTML to critical editing in TEI XML (the focus of year 1) and, beyond text and editing, discover computer-assisted analytical and visualisation methods with the *Voyant Tools* software environment and then work in a literate programming framework (For which the canonical reference is Knuth, 1984) implemented in R Markdown (the focus of year 2, see Figure 1).



```
plot(dem$Words)
```

```
plot(dem$Genre)
```

~~~
---
title: "HN3"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

```{r, include=FALSE}
getwd()
```

Cicéron

```{r Cicero, echo = TRUE}
cic_corpus <- scan("cicero_01.txt", what="character", sep = " ", quote = "")
class(cic_corpus)
str(cic_corpus)
length(cic_corpus)
mode(cic_corpus)
dim(cic_corpus)
summary(cic_corpus)
```

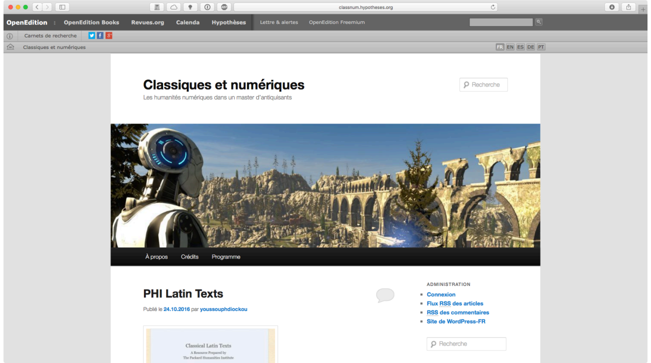
Démosthène

```{r Demosthenes, echo = TRUE}
dem <- read.csv("demosthenes_works.csv")
```

```
  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Figure 1: Text analysis in RStudio

The principles of the courses will be expounded: favouring active participation, learning-by-doing and flipped classroom teaching; insisting on the critical, reflexive dimension of digital procedures; promoting free resources like *TEI by Example* (Van den Branden, Terras, and Vanhoutte) and *The Programming Historian* (Crymble et al), as well as data reuse; developing an open publication culture through the *Classiques et numériques blog* maintained by the students (see Figure 2) or a shared Zotero group library; creating an awareness of the surrounding Digital Humanities communities; fostering actual collaboration, both between the students and with other projects or programmes – to date, with another MA specialised in Web design on an online edition prototype, with the Pelagios Commons project on the annotation of place names and with the Sunoikisis Digital Classics network in its effort to collectively define a core syllabus.



Classiques et numériques

Les humanités numériques dans un master d'antiquisants

PHI Latin Texts

Publié le 24.10.2016 par youssouphdiockou

Figure 2: Classiques et numériques, the blog of the MA

Comparing Models

Finally, drawing on this experience I will address several aspects of the current development of Di-gital Humanities pedagogy: as a separate entreprise or within established disciplines, with or without infrastructural, collegial or cross-departmental support, in various time formats, with different modes of external collaboration, etc. To sketch this broader typology, I will compare this French series of courses with other models, using in particular the data contributed to the aforementioned Digital Humanities Course Registry.

The poster will be in English, but I will naturally interact with the audience of the poster sessions both in English and French.

Bibliography

Crymble, A., Gibbs, F., Hegel, A., McDaniel, C., Milligan, I., Taparata, E., Visconti, A. and Wieringa, J. (eds.) (n.d.) . *The Programming Historian*. http://programminghistorian.org/.

Hirsch, B. (ed.) (2012). *Digital Humanities Pedagogy: Practices, Principles and Politics.* Open Book Publishers. http://www.openbookpublishers.com/reader/161.

Jockers, M. (2014). *Text Analysis with R for Students of Literature.* New York: Springer.

Knuth, D. (1984). "Literate Programming." *The Computer Journal*, 27(2): 97–111. http://comjnl.oxfordjournals.org/content/27/2/97.short.

Rockwell, G. and Sinclair, S. (1984). *Hermeneutica. Computer-Assisted Interpretation in the Humanities*. Cambridge, Massachusetts: MIT Press.

Sahle, P. (2013). "DH Studieren! Auf Dem Weg Zu Einem Kern- Und Referenzcurriculum Der Digital Humanities." *DARIAH-DE Working Papers,* 1. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2013-1.pdf.

Van den Branden, R., Terras, M. and Vanhoutte, E. (n.d.) "TEI by Example." accessed 1 November 2016. http://teibyexample.org/.

Campus « Archives Audiovisuelles de la Recherche »

Christine Berthaud
christine.berthaud@ccsd.cnrs.fr
CNRS, CCSD, France

Peter Stockinger
peter.stockinger@inalco.fr
ESCoM-FMSH, INALCO, France

Valérie Legrand
vlegrand14@yahoo.com
ESCoM-FMSH, France

Steffen Lalande
slalande@ina.fr
INA, France

Abdelkrim Beloued
abeloued@ina.fr
INA, France

Rania Soussi
rania@armadillo.fr
ARMADILLO, France

Yannick Barborini
yannick.barborini@ccsd.cnrs.fr
CNRS, CCSD, France

Campus AAR est un projet qui vise à développer une plate-forme scientifique pour la production, la description et la publication d'archives audiovisuelles numériques pour la recherche, l'enseignement et la culture. Ce projet a rassemblé quatre partenaires : ESCoM-AAR (FMSH) coordinateur, Armadillo, INA, CCSD, il a été financé par l'Agence Nationale de la Recherche (2014-2016)

Campus AAR s'inscrit dans le contexte des *digital humanities*, plus particulièrement dans celui de la constitution et exploitation de patrimoines scientifiques et culturels des sciences humaines et sociales (SHS) sous forme de corpus ou d'archives audiovisuelles en ligne en vue de leur usage dans la recherche, l'enseignement et la communication-valorisation.

Cette approche correspond aux attentes et besoins des principales parties prenantes d'archives de recherche (i.e. enseignants ; chercheurs, étudiants, doctorants, ...) qui souhaitent intervenir *activement* sur un média audiovisuel (une vidéo, une image) pour le *transformer* en une *ressource intellectuelle à proprement parler* qui leur sert dans leurs activités spécifiques d'apprentissage, d'enseignement, de recherche, etc.

En adoptant une approche cognitive et sémiotique des archives audiovisuelles, le projet Campus AAR se propose de développer une plateforme scientifique et technique pour l'archivage, la description et la publication de vidéos scientifiques diffusées en accès ouvert sous forme d'analyses audiovisuelles, de corpus enrichis et de portails web.

L'objectif principal de Campus-AAR est de mettre au point une infrastructure logicielle déployable ainsi qu'un ensemble de ressources sémantiques permettant d'archiver, d'analyser, d'exposer, de publier, de valoriser, de rechercher et de rendre interopérables des ressources audiovisuelles selon des approches scientifiques variées.

Ce projet a rassemblé quatre partenaires : ESCoM-AAR (FMSH) coordinateur, Armadillo, INA, CCSD) il a été financé par l'Agence Nationale de la Recherche (2014-2016)

Méthodologie et Résultats

Les partenaires ont mis au point une méthodologie de R&D, reposant sur un travail d'expérimentation des outils, mené en étroite concertation avec les acteurs impliqués dans les Humanités Numériques.

La réalisation de scénarii d'usage, correspondant à des cas d'application typiques, a permis de développer et d'évaluer l'environnement Campus-AAR.

Des phases de tests, retraçant les principales étapes de travail avec Campus-AAR, ont été effectuées par un panel représentatif de professionnels, correspondant aux profils types des utilisateurs-cibles du projet.

A l'issue de ces phases de développement, le projet Campus-AAR a réalisé :

- La plateforme d'éditorialisation Campus AAR: une solution logicielle "communautaire" pour l'archivage, l'analyse et la publication de ressources AV
- Les sites et portails audiovisuels Campus AAR : le site central, les collections vidéos sur MediHAL et les portails thématiques constitués d'analyses audiovisuelles (AGORA: Patrimoine en SHS, ARC: Diversité culturelle, AHM: Histoire des Mathématiques).

Le projet Campus-AAR s'adresse aux acteurs du monde scientifique, culturel et universitaire (chercheurs, enseignants, archives, laboratoires, musées, etc.) souhaitant travailler sur un média audiovisuel pour le transformer en une ressource intellectuelle adaptée à des contextes d'usage variés. Il prévoit de mettre à disposition des utilisateurs une plateforme communautaire pour l'archivage, l'analyse et la publication de ressources audiovisuelles, réunissant :

- Le portail central : accès aux logiciels, à la documentation, aux collections et portails audiovisuels
- Les Archives MediHAL - Campus AAR : espace d'archivage pérenne et de diffusion des vidéos - *CCSD*
- Le Studio Campus AAR : logiciel d'édition d'ontologies, d'analyse et de publication audiovisuelles - *INA*
- Le CMS Campus AAR : interface de publication des analyses AV et de gestion des portails - *Armadillo*
- Le Centre de Ressources Métalinguistiques Campus AAR: ressources conceptuelles communes : modèles de données, ontologies, thesaurus - *ES-CoM-AAR*
- Le Back Office triplestore Campus AAR : un environnement pour le traitement des données (gestion des utilisateurs, moteur de recherche, etc.)

Big-Data Oriented Text Analysis for the Humanities: Pedagogical Use of the HathiTrust+Bookworm Tool

Sayan Bhattacharyya
sayanb@sas.upenn.edu
University of Pennsylvania, United States of America

Christi Merrill
merrillc@umich.edu
University of Michigan, United States of America

Peter Organisciak
organis2@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Benjamin Schmidt
b.schmidt@northeastern.edu
Northeastern University, United States of America

Loretta Auvil
lauvil@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

Erez Lieberman Aiden
erez@erez.com
Rice University, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois – Urbana-Champaign
United States of America

The HathiTrust Bookworm (HT+Bookworm) is an interactive tool for visualizing content from the HathiTrust Digital Library, which contains almost 15 million volumes of digitized text (Auvil et al. 2015). Our poster describes the application of HT+Bookworm in teaching. Studies show that the complexity of integrating into pedagogical practice text analysis tools operating over large datasets is a significant barrier to their uptake (Green et al. 2016), and that the best solution is to use a pre-populated text analysis tool (Sinclair and Rockwell 2012, Rockwell et al. 2010). HT+Bookworm accomplishes this by facilitating exploration of the "big" textual data of the HathiTrust Digital Library's collection without requiring mastery of complex technology. We recently used HT+Bookworm in class sessions co-taught by two of us as part of literature classes at the University of Michigan, Ann Arbor to undergraduate students who had no prior familiarity with quantitative approaches to text analysis. One of our objectives was to help students discover how the meanings of words can vary — both when word meanings change over time, and when the same word, when borrowed from one discipline or domain and applied to a different discipline or domain (or simply applied independently in two different domains), takes on separate meanings.

Rens Bod argues that it is only when humanistic disciplines are compared on a large scale that patterns across them become visible (Bod 2013). HT+Bookworm enables the discovery of such patterns by facilitating comparison across categories of knowledge. While search engines are adept at finding individual texts within a digital library, Bookworm performs a different task — abstracting across categories within the library and visualizing those abstractions. These abstractions, which are a form of 'distant reading' (Moretti 2013), are generated in the context of a specific textual fragment (for example, a word or phrase), through the variation of some attribute of the manifestation of that text fragment across the categories defined by some categorization scheme. A typical categorization scheme is the organization of the digital library by Library of Congress (LoC) classes as metadata, and a typical attribute of the manifestation of a word or phrase across categories is its normalized frequency of occurrence across those categories. HT+Bookworm works with both discrete sets of categories such as LoC classes as well as with continuous categories such as time. A time-series plot of the normalized frequency of occurrence of a word over a chronological range, within certain specified LoC classes, provides a sense of how the relative occurrence of that word or phrase has varied across those LoC classes over the specified time range. Students generate visualizations consisting of layered time-series plots (stacked area charts) for the relative frequency of their words of interest, within determinate categories of interest in the HathiTrust Digital Library collection. These categories correspond to the layers (stacks) of the plot, with each layer encompassing an LoC class and time represented along the x-axis. The HT+Bookworm tool also provides, at each point in the plot, a subset list of volumes that contribute the most to the attribute being plotted. This list, accessible by mouse-click at the requisite point, serves as the gateway to the digitized text of the individual volumes in the list. This affordance helps students bridge the gap between the abstraction of distant reading and the discovery of specific texts that they can then investigate further through close reading.

Our poster includes instances of the kinds of exploration HT+Bookworm made possible for students. An example follows. The concept of "fidelity" (an important word to explore in connection with translation studies, a topic of the classes) shows different characteristics when explored in English (in which the concept maps onto the two words "fidelity" and "faithfulness") and in Spanish (where the concept maps onto the single word "fidelidad"). Investigating the occurrence of the word by LoC category allows students to explore hypotheses such as whether the greater strength, historically speaking, of religious tradition in the Spanish-speaking world in comparison with the Anglophone world affects the relative prevalence of this word in different domains of use. Another example is a stacked area chart for a word of a kind for which HT+Bookworm helps provide an understanding of the word's differentiated meanings in different use categories (for example, in the case of the word "depression", in the use category of psychology and medicine versus that of economics). HT+Bookworm accomplishes this by abstracting separately across different LoC classes, while aiding in the indication of the points in time at which, for each class, the word entered widespread usage.

Bibliography

Auvil, L., Aiden, E. L., Downie, J.S., Schmidt, B., Bhattacharyya, S. and Organisciak, P. (2015). "Exploration of Billions of Words of the HathiTrust Corpus with Bookworm: HathiTrust + Bookworm Project." Digital Humanities 2015 (DH 2015) Conference, Sydney, Australia. 29 June - 3 July 2015.

Bod, R. (2013). A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present. New York: Oxford University Press.

Green, H., Dickson, E. and Bhattacharyya, S. (2016). "Scholarly Requirements for Large Scale Text Analysis: A User Needs Assessment by the HathiTrust Research Center." Digital Humanities 2016 (DH 2016) Conference, Krakow, Poland. July 2016.

Moretti, F. (2013). Distant Reading. London: Verso.

Sinclair, S. and Rockwell, G. (2012). "Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies." In Brett D. Hirsch (ed.), Digital Humanities Pedagogy: Practices, Principles and Politics. Cambridge, U.K.: OpenBook Publishers, pp. 241-64.

Rockwell, G, Sinclair, S., Ruecker, S. and Organisciak, P. (2010). "Ubiquitous Text Analysis." paj: The Journal of the Initiative for Digital Humanities, Media, and Culture, Vol. 2, No. 1.

Database of Belarusian Literary Periodicals

Ingo Börner
ingo.boerner@univie.ac.at
University of Vienna, Austria

Gun-Britt Kohler
gun.b.kohler@uni-oldenburg.de
University of Oldenburg, Germany

Sünna Looschen
s.looschen@uni-oldenburg.de
University of Oldenburg, Germany

Introduction

This paper presents the prototype of a "Database of Belarusian Literary Periodicals" developed at the University of Oldenburg, Germany. The project addresses the hitherto under-researched area of Belarusian literature from a field theoretical perspective (Bourdieu, 2002) in combination with a quantitative approach to Literary History (Moretti, 2005).

Belarusian Literature

The case of Belarusian literature is interesting in particular because it constitutes a highly unstable literary field. Due to historical reasons the formation of the literary field and the rise of a literary market dates back only to the beginning of the 20th century. During the first half of the 1920s the establishment of proletarian society in Soviet Belarus entailed the promotion of Belarusian language and literature, which were then consolidated by specific institutions. But already from the second half of the 1920s, cultural politics changed and literary life came under increasing ideological control ending in the infamous 'cleansings' ("čistki") of the 1930s.

Literary Periodicals

Research on the formation of literary markets and fields has illustrated the crucial role of literary periodicals (Bourdieu, 2002; van Rees, 2012). We consider these magazines media that allows us to reconstruct and to analyze the specific structure and internal development of the Belarusian literary field. They allow us to trace the configurations of authors within groups and magazines, the trajectories of single authors and/or literary groups in the field, the differentiation of the literary genre system, the formation of literary criticism and so forth.

The Database

The prototype of the database includes the four most important Belarusian literary periodicals published between 1922 and 1939: *Maladnjak*, *Polymja*, *Uzvišša* and *Kalos'se* (cf. Kohler, 2016). For the time being, we focus on capturing these periodicals' tables of contents, assuming that a systematic analysis of literary periodicals does not necessarily require the literary texts themselves but that the corresponding 'paratexts' suffice (cf. Genette, 1989).

The database currently comprises the tables of contents of 252 issues that were transcribed manually and were encoded according to the TEI-Guidelines. The database itself is set up as an application for the open-source XML database eXistdb. An extensive range of search queries can be performed on the corpus, which enables users to identify quantitative characteristics of the periodicals and its contributing authors. It allows for the export of the data as a dynamic network graph in the gefx-format (Gephi).

Pilot study: corpus, questions, and method

For the pilot study we specifically focus a) on the period 1922–1932 (the year 1932 marks the end of literary diversity: groups and magazines dissolved and fused into the 'one' Writers' Association with its official periodical *Polymja*), and b) on the three periodicals in Soviet Belarus (*Kalos'se* was established only in 1934 and was published in the Polish part of the country). Taking into account this focus, the pilot study deals with a complete corpus of 189 issues.

We focused on the analysis of c) authors' trajectories (fluctuations between periodicals), including the question of "splinter groups' trajectories" and of d) hierarchization of authors. These questions are interlinked with the questions, e) whether an author's movement from one periodical to another brings an increase of his/her publication frequency, and how the constellation of authors changed in relation to such movements.

Step 1: Identification of authors: Problems we had to overcome in this respect were variations in spelling and the frequent use of acronyms and pseudonyms. The use of pseudonyms is in some cases not only linked to the periodical (some authors published in one magazine using their real name and under pseudonyms in the others) but also to the author's role (e.g. writer vs. critic/reviewer).

Step 2: Identified authors were linked to a <person> record, where additional information on each identified author is stored. If available, these entries were linked to authority files (VIAF). As the coverage of Belarusian authors in the relevant authority files is rather incomplete, the authors were also linked to the corresponding Wikidata entries, to provide for an external unique identifier.

Step 3 (question 'd'): We identified frequently published authors and compared their rankings in the periodicals at several points in time but also focused on literary reviews and translations.

Step 4 (questions ‘c’, ‘e’, ‘f’): We observed a relatively high fluctuation of authors and splinter groups between the periodicals, including also the authors that published most frequently. In analyzing the movements of authors, it was possible to complement hitherto lacking insights and background knowledge about the journals’ interconnectedness.

Bibliography

Bourdieu, P. (2002). *The Rules of Art: Genesis and Structure of the Literary Field*. Cambridge: Polity Press.

Genette, G. (1989). *Paratexte.* Frankfurt a. Main.

Kohler, G.-B. (2016). “‘Success’ and ‘Failure’ of Literary Collaboration between Authors in Belarus in the 1920s.” In Butler, M., Hausman, A. and Kirchhofer, A. (eds), *Precarious Alliances. Cultures of Participation in Print and Other Media.* transcript, pp. 207–40.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London – New York: Verso.

Van Rees, K. (2012). “Field, Capital and Habitus. A Relational Approach to ‘Small’ Literatures.” In Kohler, G.-B., Navumenka, P. and Grüttemeier, R. (eds), *Kleinheit als Spezifik. Beiträge zu einer feldtheoretischen Analyse der belarussischen Literatur im Kontext ‘kleiner’ slavischer Literaturen.* Oldenburg, pp. 5–56.

DH Uncentered: Fostering Partnerships Within a Digital Humanities Network

Laura Braunstein
lrb@dartmouth.edu
Dartmouth College, United States of America

Scott S. Millspaugh
scott.s.millspaugh@dartmouth.edu
Dartmouth College, United States of America

During the academic year 2014-15, Dartmouth College launched an informal Digital Humanities initiative. Individual researchers and campus stakeholders already constituted a loose digital humanities network at the college, and had for years, but these projects had not been coordinated through any kind of centralized infrastructure. In spring 2015, Laura Braunstein (Digital Humanities Librarian) and Scott Millspaugh (Instructional Designer), proposed to pilot a faculty residency program during the summer term. Funded by the Dartmouth College Library and Information Technology Services (ITS), the residency program had the following objectives:

- to prototype a team-based approach to supporting digital humanities projects that could be standardized for future implementation; and
- to test the capacity of the Library and ITS to support digital humanities projects without recourse to external sources of funding.

The project selected for the Residency Program Pilot was *Multimedia in the Long Eighteenth Century* (MMLEC), led by faculty PI Scott Sanders (Assistant Professor of French). The MMLEC project seeks to quantify the frequency with which musical paratext, including both lyrics and musical notation, appear in English- and French-language novels published between 1688 and 1815. Professor Sanders, John Wallace (Research Systems Engineer), and Jill Baron (Romance Languages Librarian) had been collaborating on the project since late 2014, and Sanders and Wallace had already presented a plan for the project at the Digital Libraries for Musicology Workshop 2015. The Residency Program Pilot allowed for the expansion of the core project team and funded the employment of two student workers who created a trained data set under the direction of Mark Boettcher (Senior Programmer and Analyst). The Pilot also led to the involvement of the Digital Library Technologies Group (DLTG), which has assisted in the creation of a bibliographic database, still in development, of published French-language works from the eighteenth-century.

The Residency Program Pilot was successful in establishing a model for institutional support of other DH projects in the pipeline at Dartmouth and has allowed the Library and ITS to develop a process for evaluating and prioritizing those projects. In the summer of 2016, we renewed the Residency Program for another year with the Jamaican Slave Names Project (JSNP). Our poster will narrativize the development of MMLEC and JSNP within the context of the Residency Program and demonstrate how institutional support of a digital humanities project can help foster community and cross-campus collaboration in the absence of a dedicated DH program or center. We hope to dialogue with colleagues at other international institutions present at DH2017 that have established different models of support for digital projects in the arts and humanities and to offer guidance to those that are initiating cross-campus collaborations within a network of individual researchers, centers, and administrative groups, instead of concentrating digital humanities support within a single program or center.

Humanities Data Centre (HDC) – Developing Services for Heterogenous Humanities Research Data

Stephan Buddenbohm
stefan.buddenbohm@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Claudia Engelhardt
claudia.engelhardt@sub.uni-goettingen.de
Göttingen State and University Library, Germany

Sven Bingert
sven.bingert@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH
Göttingen, Germany

Research data play a fundamental role in science and the arts and humanities, a large part of them being either digitised or of digital provenance nowadays. Digital research data are fragile, which means that their thorough management and preservation is crucial to prevent the loss of data and information, avoid redundant data collection, enable more efficient research, ensure the verifiability and reproducibility of research results as well as the citability and reusability of research data.

The management and long-term preservation of digital research data requires an infrastructure that takes into account the digital nature and heterogeneity of the data and the diverse requirements of the research community. Research data centres, such as the Humanities Data Centre (HDC), are key players in this context. They provide a safe and trustworthy place for researchers to deposit their data as well as to search for and get access to previously deposited data. They also act as centres of expertise, offering training and support to their community. The importance of case-specific consultation must not be underestimated as the heterogeneity of digital research data and the wide range of data management tasks across the data life cycle call for extensive support by data experts.

Comprising of a large number of disciplines and employing a variety of different research methods, the arts and humanities produce very heterogeneous research data. A great deal of them are file-based data, such as text, audio or video files. But there is an increasing number of more complex research data, such as digital editions or visualisation frameworks, that consist of various interwoven layers of different types of data. In these instances, it is often hard or impossible to distinguish between the "primary data" and the software application that is used to process and display the data. While there are more or less established solutions for file-based data (e.g. repositories), the curation and long-term preservation of complex data types still poses a challenge to infrastructure providers.

In view of this background, the HDC developed an initial service portfolio in its design phase that responds to the heterogeneity and complexity of arts and humanities research data by offering consultation and training, repository services for file-based data and the application preservation. The latter is designed to sustain representations of complex research data such as digital editions or visualisation frameworks.

What are the challenges for the research data centre in this? Continuous access to static systems in a developing environment will create problems sooner or later - the systems will eventually become outdated and inaccessible. On the technological level, there are basically two concepts to preserve applications: either virtualisation or emulation of components. While emulation allows the operation of software (such as an operating system) on hardware and software environments it had not been developed for, it costs extra computing power to emulate the required environment, to name only one disadvantage. Virtualisation comes with the clear benefit that the hardware appears as a physical device to the system and to the software modules. The disadvantage is that the virtualised environment must be sufficiently performant directly on the hardware. But for most of the use cases considered here, only standard hardware is being used. Hence the advantage of running more than one virtualised system and the relatively simple management of these systems makes the virtualisation the state of the art approach. The poster illustrates how the application preservation of the HDC uses virtualisation to preserve complex research data applications in a protected environment and ensure their representation, citation and use for a finite period of time.

Mapping Linked Data Subject Headings in the Library Catalog

Patrick Burns
patrick@diyclassics.org
Institute for the Study of the Ancient World Library United States of America

Gabriel McKee
gm95@nyu.edu
Institute for the Study of the Ancient World Library United States of America

David M. Ratzan
dr128@nyu.edu
Institute for the Study of the Ancient World Library United States of America

Tom Elliott
te20@nyu.edu
Institute for the Study of the Ancient World
United States of America

The MARC-based library catalog is text-based: in order to find relevant information, a researcher must use text to identify an author, title, or subject. Our team has been expanding the paradigm of text-based discovery by exploring visual discovery, building a browsable map of our library catalog based on record authority and linked open data. This geospatial representation of our holdings not only offers researchers a new mode of discovery; it also opens the door for new avenues of research by highlighting unexpected connections and virtually collocating materials that are classified and shelved separately. The process is as follows. 1. Newly catalogued items are assigned a stable URI from Pleiades, the open-access ancient world gazetteer (Bagnall et al., 2016), reflecting our Institute's scholarly focus and library holdings. 2. This identifier is added as a subject heading in the item's MARC record. 3. This data is exported, cross-referenced with geoJSON records containing latitude and longitude data (Baumann, 2014), and mapped using Leaflet.js (Agafonkin, 2016). The mapped representation of the catalog serves as an alternative mode of discovery for researchers, who can now browse for library materials by focusing on, for example, a general region of the Near East, a city in ancient Egypt, or a specific archaeological site in China. The experimental nature of this map-based discovery has become all the more viable as the Library of Congress has recently added Pleiades to its list of subject heading authorities. (Library of Congress, 2016) So, while the inclusion of linked data in MARC records has seen some adoption in recent years (Papadakis et al., 2015), this imprimatur from the LOC for Pleiades opens up massive potential for geographical linked data specifically. Other projects have mapped LOC subject headings (e.g. Freeland et al., 2008; Bennett et al., 2011), but this is the first project to do so directly from linked data embedded in MARC records. Accordingly, this project heeds the recent call of the MIT Future of Libraries report (MIT Libraries, 2016) to provide "comprehensive, accessible, digital content" in library discovery that supports the ability to "combine, manipulate, [and] visualize" library data for the global community. Our poster includes: an explanation of the linked-data principles underlying the project, our visualization workflow, and an example of mapped catalog data from books acquired in the fourth quarter of 2016, presented both as a static map on the poster as well as a live, browsable demo on a tablet.

Bibliography

Agafonkin, V. (2016). Leaflet : An Open-Source JavaScript Library for Interactive Maps. http://leafletjs.com/ (accessed 28 March 2017).

Bagnall, R.S., et al. (2016). Pleiades: A Gazetteer of Past Places. http://pleiades.stoa.org/ (accessed 28 March 2017).

Baumann, R. (2016). Pleiades-Geojson. https://github.com/ryanfb/pleiades-geojson (accessed 28 March 2017).

Bennett, R., O'Neill, E.T., Kammerer, K., and Shipengrover, J. D. (2011). mapFAST: A FAST Geographic Authorities Mashup with Google Maps, The Code4Lib Journal, 14. http://journal.code4lib.org/articles/5645 (accessed 28 March 2017).

Freeland, C., Kalfatovic, M., Paige, J., and Crozier, M. (2008). Geocoding LCSH in the Biodiversity Heritage Library, The Code4Lib Journal, 2. http://journal.code4lib.org/articles/52 (accessed 28 March 2017).

Library of Congress, Network Development and MARC Standards Office. (2016). Technical Notice September 15, 2016: Additions to the MARC Code Lists for Relators, Sources, Description Conventions. https://www.loc.gov/marc/relators/tn160914src.html (accessed 28 March 2017).

MIT Libraries, Ad Hoc Task Force on the Future of Libraries. (2016). Institute-Wide Task Force on the Future of Libraries—Preliminary Report https://future-of-libraries.mit.edu/sites/default/files/FutureLibraries-PrelimReport-Final.pdf (accessed 28 March 2017).

Papadakis, I., Kyprianos, K., and Stefanidakis, M. (2015). Linked Data URIs and Libraries: The Story So Far, D-Lib Magazine, 21(5/6). http://www.dlib.org/dlib/may15/papadakis/05papadakis.html (accessed 28 March 2017).

A New and Improved Method to Text-Mining in Chinese: Closer Language Segmentation in Detecting the Shifting Meaning of Patriotism

Annie S. Chao
mrsannechao@gmail.com
Rice University, United States of America

Qiwei Li
liqiwei2000@gmail.com
Rice University, United States of America

We aim to demonstrate a new methodology for detecting shifting nuances of ideas in Chinese intellectual history. Our subject is Chen Duxiu (1879-1942), founder of the Chinese Communist Party (CCP) and one of the most important historical figures of twentieth century China. We combine the latest text-mining tools with statistical analysis and natural language law, based on word frequency calculations. Because the Chinese language does not have spaces between words, and because each word is comprised of either a single or multiple characters or morphemes, segmentation of Chinese text poses a set of different challenges than for English corpus. By using a Chinese tokenization plug-in for R called JiebaR, we have developed a more precise method to create a Chinese natural language curve that conforms closely to Zipf's law, a commonly used model for distribution of words in a corpus. Zipf's law states that given a body of natural language text, the frequency of any word is inversely proportional to its rank. (Ha et al., 2003) For Chinese language, Zipf's law applies for words made up of multiple morphemes. (Xiao, 2008)

Our assumption is that a word is significant when its position on our curve deviates from the fitted curve based on Zipf's law. Departing from an existing method developed by Prof. Jin's team, we created two groups of keywords, and called them "anchor" and "companion" words. (Jin et al., 2014) "Anchor" words have large residuals (high deviation from the standard curve), and "companion" words have a high correlation with "anchor" words. We used the formula for Pearson's correlation coefficient to find the companion words. The coefficient has a value of between +1 and -1. The companion words provide the context with which to interpret the anchor words. Unlike Prof. Jin's team, we included keywords made up of more than two characters, such as the three character word for Nationalist party, Guomindang (國民黨), the four character word for citizen's assembly, guomin huiyi(國民會議), and the five character word for Marxism, makesizhuyi (馬克思主義).

The goal of our research is to analyze changing meaning of the concept of patriotism in Chen's writing from the beginning of his publishing career, ca. 1897, to the end of his life, 1942. Why is Chen such a fascinating figure to study? In addition to creating a political party that changed the course of Chinese history, Chen was equally, if not more, influential in bringing about the first cultural revolution of twentieth century China. Living at a time of great political unrest, Chen wrote passionately on the need to reform the people by revolutionizing Chinese culture, thoughts, and politics. After founding the CCP in 1920-21, Chen was expelled from the party in 1929 due to ideological differences. He was subsequently jailed by Chiang Kai-shek's Nationalist Party, and died a political pariah a few years later. As his political and personal fortune waxed and waned, Chen's conception of patriotism shifted. For this paper, we chose six important anchor words: citizen (國民), youth (青年), democracy (民主), revolution (革命), people (民族) and being patriotic (愛國). Our method is replicable for other corpus, with the understanding that the conclusion is derived with the additional layer of human deliberation.

Bibliography

Ha, L.Q., Sicilia-Garcia, E.I., Ming, J., and Smith, F.J. (2003). Extension of Zipf's Law to Word and Character N-grams for English and Chinese, *Computational Linguistics and Chinese Language Processing,* 8:1, 77-102.

Jin, G., Leong, Y., Yu, Y., and Liu, C., (2014). Application of Statistical Residual Analysis to Humanities Studies: Using Xin Qing Nian as Example, *Journal of the History of Ideas in East Asia*, 6: 327-366.

Xiao, H., (2008). On the Applicability of Zipf's Law in Chinese Word Frequency Distribution, *Journal of Chinese Language and Computing*, 18:1, 33-46.

The new literacy practice of young Taiwanese writers, illustrators, social innovators, and makers

Su-Yen Chen
suychen@mx.nthu.edu.tw
National Tsing Hua University, Taiwan

Jason S. Chang
jschang@cs.nthu.edu.tw
National Tsing Hua University, Taiwan

Hsing-Yu Chang
hsich179@gmail.com
National Tsing Hua University, Taiwan

Hsin-Yu Kuo
kuohy@mx.nthu.edu.tw
National Tsing Hua University in Taiwan, Taiwan

Yu-Hsuan Wu
shanny@nlplab.cc
National Tsing Hua University, Taiwan

New literacy is not to be taken as merely a change in how we understand and present information because of newly-developed technologies, but as a way of transforming how we interact with texts and with the world as a whole. This study aims to report an empirical study on how four types of digital natives, all born in the 1980s, access information and create content. Data was collected through one-on-one semi-structured interviews with 28 writers, illustrators, social innovators, and makers in total, seven in each category. The data were then processed as follows: First, we employed qualitative analysis and interpretation through close readings and meaning-making. Second, we quantitatively deconstructed the text by atomizing the discourse with Chinese Knowledge and Information Processing (CKIP) and filtering out units that are irrelevant to the analysis. Finally, we generated comparisons among the four contemporary groups and their new literacy practices.

Writers

In contrast to their counterparts from the older generation who were devoted to a single identity such as being a novelist, columnist, or essayist, young creative writers who have won literature awards and published top-ranking books grew up within a social environment of PTT, blogs, and Facebook. They are accustomed to managing their primary literate role while at the same time enjoying multitasking by being active across new media platforms and acting as popular commentators for social issues related to their academic training such as education, sociology, gender study, and philosophy. They obtain a vast quantity of high quality information from their social media stratosphere and they share opinions and ideas through social media services with their followers as a form of self-presentation, or “performance” in Goffman’s term.

Illustrators

The internet and social media have brought changes to people’s reading habits in many ways; for example, “Picture Reading” and illustrated creative blogs have become very popular in Taiwan over the last decade. Therefore, illustrators who devote effort to character creation and social media fan page management are increasing in number. They obtain inspiration from daily life and can transform their observations into creative content. These illustrators integrate their interests and capacities to share their artwork on the Internet, and they attempt to develop a business model that involves a multivariate form of production.

Social innovators

Social innovators attempt to employ strategies and actions to solve community or city problems through teamwork with the goal of making the world a better place. They use technology as a means of connecting resources in the process. Their goal is to penetrate social problems and issues, identify the needs of communities they target, develop innovative solutions, and take action. One of their primary strategies for achieving this goal is to collaborate with local people to create stories for the public, who thus can understand the value and meaning behind these stories. Their storytelling is multimodal in nature to accommodate the needs of the general public.

Makers

The “maker movement” has emerged in Taiwan over the last several years, and under the social environment of learning by doing, makers create a small amount of delicate products and are very open regarding sharing their ideas and creative processes. They suggest that traditional education in Taiwan neglects the significance of DIY and are enthusiastic about the culture of makers, who emphasize learning, doing, and sharing. They are very persistent about their own interests, and empower themselves with knowledge and skills through self-directed learning, pursuing an ideal living style and working arrangement.

Insight from word clouds and the co–occurrence of key words

The word clouds of writers illuminate a central focus is on words, articles and literature. The literacy practice of these writers is to express feelings and provide arguments about issues. By contrast, illustrators are interest-driven; they enjoy sharing life stories and ideas with their friends, and are followed by social media fans who adore the characters they have created. The central focus of social innovators is on society, community, and local issues and needs; and their resources and contributions include teamwork, space, activities, cultural artifacts, and communication between people and the government. Finally, makers are interest-driven and they enjoy sharing life experiences through their creations and products like illustrators. Much like social innovators, makers access information and create content in both the cyber world and in the physical world.

La localisation du jaune dans des dessins de dieux réalisés par des enfants

Christelle Cocco
christelle.cocco@unil.ch
University of Lausanne, Switzerland

Damien Firmenich
damien.firmenich@epfl.ch
University of Lausanne, Switzerland

Pierre-Yves Brandt
pierre-yves.brandt@unil.ch
University of Lausanne, Switzerland

Sabine Süsstrunk
sabine.susstrunk@epfl.ch
University of Lausanne, Switzerland

Dans le cadre de la recherche interdisciplinaire, “Drawings of gods” (financé par le Fonds National Suisse), ancrée en psychologie de la religion et visant à comprendre les stratégies cognitives mises en œuvre par les enfants pour dessiner “dieu” (Brandt et al., 2009, Dandarova, 2013, Brandt, 2016), une question s’est posée quant à l’utilisation du jaune : “Est-ce que le jaune est une couleur privilégiée dans la représentation de dieux parce que la lumière est souvent associée au divin et que le jaune est utilisé pour représenter la lumière ?”.

Afin de répondre à cette question, une première annotation manuelle des dessins collectés par les chercheurs dans différents pays a été effectuée dans une feuille de calcul, spécifiant dans quelle zone de l'image se trouve le jaune (au milieu, autour de la figure principale de dieu, en périphérie), s'il y avait présence d'un soleil jaune ou pas de jaune.

Ce projet comptant actuellement plus de 6'500 dessins, il est devenu nécessaire de faire appel à une annotation automatique. Aussi, il est intéressant, du point de vue de la psychologie, de comprendre quelles décisions humaines peuvent être reproduites par un ordinateur. Bien que les méthodes de traitement d'images et de vision par ordinateur, combinées aux algorithmes de classification supervisée, soient très développées et performantes pour le traitement d'images naturelles (Szeliski, 2010), elles le sont beaucoup moins pour le traitement de dessins (Stork, 2009). Par conséquent, ce travail consiste à explorer les techniques possibles et à trouver des caractéristiques ("features") pertinentes pour une classification supervisée.

Pratiquement, après avoir défini ce qu'était la couleur jaune dans les dessins et extrait cette dernière, toutes les images ont été transformées en format carré, afin de permettre la comparaison des dessins au format paysage et au format portrait. Ensuite, différentes approches ont été testées pour reproduire l'annotation manuelle. La première consistait à extraire la gravité du jaune, définie comme la moyenne de l'intensité de cette couleur par ligne (respectivement par colonne), pour la hauteur (respectivement la largeur) de chaque dessin. Comme attendu, les courbes obtenues montrent des pics dans les zones contenant une forte concentration de jaune, clairement repérables à l'œil nu. Cependant, en raison de la variabilité de l'intensité du jaune entre les différents dessins et l'application non régulière des couleurs dans chaque dessin, il n'a pas pu être défini de critère permettant de repérer ces pics.

Dans un second temps, partant de la première approche, le centre de gravité du jaune dans chaque dessin a été extrait et deux cercles ayant pour origine ce centre ont été définis. Le but alors visé était de définir des surfaces correspondant aux zones annotées manuellement (milieu, autour ou périphérie), en se basant sur une hypothèse forte estimant que le centre de gravité du jaune correspond à celui de la figure principale. Cette méthode, qui a conduit à analyser les intensités de jaune dans les secteurs circulaires des cercles, a à nouveau été mise à mal par la variabilité dans l'ensemble des dessins.

Inspirée de l'idée de zones délimitant l'espace de la seconde approche, la dernière méthode a consisté à diviser l'image carrée en 25 cases, formant ainsi une grille de cinq par cinq. Ensuite, en fonction de l'intensité du jaune dans chacune de ces cases et en faisant l'hypothèse que la figure principale est au centre de l'image, comme souvent observé dans les dessins d'enfants (Golomb, 1987, Winner, 2006), une série de conditions ont été choisies afin de déterminer la zone dans laquelle se trouve le jaune. En raison des résultats prometteurs ainsi obtenus, une classification supervisée multi-étiquette a été effectuée, utilisant, pour chaque image, les 25 cases comme caractéristiques des données et les annotations manuelles ("ground trouth") comme étiquettes. Les meilleurs résultats ont été obtenus avec la méthode des "plus proches voisins". À nouveau, la grande variété des techniques de dessins utilisées par les enfants, tout comme la non-constance de la position de la figure principale, conduisent à plusieurs erreurs.

Ce travail, en cours, explore actuellement de nouvelles pistes, telles que la prise en compte d'autres couleurs, la modification des paramètres de l'algorithme de classification supervisée, la définition d'une nouvelle grille permettant de prendre en compte la position de la figure principale, etc. Pour conclure, il est clair que les données sous forme de dessins, souvent utilisées en psychologie, méritent une exploration systématique des méthodes permettant de les exploiter numériquement, ceci afin de pouvoir les traiter efficacement en quantité.

Bibliographie

Brandt, P.-Y. (2016). "Représentations enfantines de dieux: comparaison interculturelle." In Rainotte, G. (eds), *Qui êtes-vous pour nous apprendre nos religions?* Louvain-la-Neuve: Academia-L'Harmattan, pp. 39-59.

Brandt, P.-Y., Kagata Spitteler, Y. and Gillièron Paléologue, C. (2009). "La représentation de Dieu: Comment les enfants japonais dessinent Dieu." *Archives de Psychologie,* 74 : 171-203.

Dandarova, Z. (2013). "Le dieu des enfants : Entre l'universel et le contextuel." In Brandt, P.-Y. and Day, J. M. (eds), *Psychologie du développement religieux: Questions classiques et perspectives contemporaines.* Genève: Labor et Fides, pp. 159-187.

Golomb, C. (1987). "The development of compositional strategies in children's drawings." *Visual Arts Research*, 13(2): 42-52.

Stork, D. G. (2009). "Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature." *International Conference on Computer Analysis of Images and Patterns.* Springer, pp. 9-24.

Szeliski, R. (2010). *Computer vision: algorithms and applications.* London: Springer.

Winner, E. (2006). "Development in the arts: Drawing and music." In Damon, W., Lerner, R., Kuhn, D. and Siegler, R. (eds), *Handbook of Child Psychology, Vol 2: Cognition, Perception, and Language.* Wiley, pp. 859-904.

Digital Humanities Clinics – Leading Dutch Librarians into DH

Michiel Cock
m.p.cock@vu.nl
Vrije Universiteit Amsterdam, the Netherlands

Lotte Wilms
lotte.wilms@kb.nl
National Library of the Netherlands, the Netherlands

In 2015, an initiative was started to set up a Dutch speaking DH+Lib community in the Netherlands and Belgium, based on the example of the American communal space of librarians, archivists, LIS graduate students, and information specialists to discuss topics ‘Where the Digital Humanities and Libraries meet’. At the initial meeting it became apparent that most participants were there to learn more about digital humanities and were not (yet) in the situation where they were able to offer expertise on the subject. On the administrative level, the directors of the libraries participating in the consortium of Dutch academic libraries (UKB) also expressed the wish that librarians become more fluent in DH.

A year later, the National Library of the Netherlands (Koninklijke Bibliotheek), and the University Library of the Vrije Universiteit Amsterdam again concluded that librarians at their institutes who wanted to get involved in DH needed more training to adequately support researchers and students in this field. Therefore both institutes joined forces to develop a set of clinics on DH for librarians. The two institutes were later joined by the Leiden University Libraries. We see this as the ideal opportunity to provide these educative sessions not only to our own librarians, but also to the academic librarians of other Dutch research libraries. In essence, we want to teach our country’s librarians the ins and outs of DH in order for them to take up their natural role of facilitating and supporting research and ideally become the research partner needed in DH projects.

The aim of these clinics is to provide basic methodological competencies and technical skills in DH, for a diverse group of librarians, consisting of both subject and technical librarians with basic technical skills. The content of these sessions should enable them to provide services to researchers and students, identify remaining gaps in knowledge or skills that they could address by self-directed learning and (perhaps) to automate their daily library work. We are not setting out to turn them into programmers or data crunchers, but want to boost their knowledge level to where they feel comfortable providing information about DH projects, follow the literature and research, follow online tutorials and hopefully take up the challenge of finishing this professional development by engaging with the DH community.

In order to design this curriculum we follow a four step approach with a Working Out Loud-principle (Williams, 2010):

1. Desk research about what being a DH librarian entails (e.g. Hartsell-Gundy et al., 2015; Mulligan, 2016; also see the Zotero library of the LIBER Digital Humanities working group);
2. Identify possible subjects, based on experience, a comparison of existing teaching material related to DH (e.g. The Programming Historian, the Digital Scholarship Training Programme at the British Library and Columbia University's Developing Librarian project) and the TaDiRAH taxonomy of research activities;
3. Get feedback from researchers on possible subjects, based on the knowledge and skills they feel librarians need;
4. Get feedback from librarians on possible subjects, based on already known gaps in their knowledge and skills.

With these in hand, we will design the curriculum of clinics, based on the method of 'constructive alignment' (Biggs et al., 2011), to make sure that the intended learning objectives and the teaching/learning activities stay aligned.

Our plan is to organize a maximum of 6 clinics, each one full day. Each day starts with one or more lectures by researchers, that address the conceptual knowledge needed. The afternoon sessions will be devoted to the hands-on training of skills, following the Library Carpentry model as closely as possible. By having researchers provide the lecture sessions, we hope to fuel the enthusiasm of the librarians with the inspiration of direct contact with researchers and to provide access to a network within and across universities. With these clinics, we hope to initiate a stream of DH activities in Dutch universities, making access to support easier for new digital scholars.

The poster at DH2017 will present the curriculum, its position in the international context and offer the lessons learned from both the design process and the first clinics. We welcome discussion about our efforts and the possibilities of applying this in other contexts.

Bibliography

ACRL Digital Humanities Interest Group. (2012-) *dh+lib* . http://acrl.ala.org/dh/ (accessed 31 March 2017g).

Baker, J. (2014) British Library Digital Scholarship Training Programme: a round-up of resources you can use - *Digital scholarship blog*. 30 October 2014. http://blogs.bl.uk/digital-scholarship/2014/10/british-library-digital-scholarship-training-programme-round-up-of-resources-you-can-use.html (accessed 31 March 2017c).

Biggs, J. B., Tang, C. and Society for Research into Higher Education (2011). *Teaching for Quality Learning at University: What the Student Does*. Philadelphia, Pa.]; Maidenhead, Berkshire, England; New York: McGraw-Hill/Society for Research into Higher Education : Open University Press.

DARIAH (n.d.) TaDiRAH - Taxonomy of Digital Research Activities in the Humanities http://tadirah.dariah.eu/vocab/index.php (accessed 31 March 2017)

Hartsell-Gundy, A., Braunstein, L. and Golomb, L. (2015). *Digital Humanities in the Library: Challenges and Opportunities* *for Subject Specialists*. Chicago: Association of College and Research Libraries, a division of the American Library Association.

Library Carpentry. (n.d.) Library Carpentry – Software skills for library professionals http://librarycarpentry.github.io/ (accessed 31 March 2017d).

Mulligan, R. (2016). *SPEC Kit 350: Supporting Digital Scholarship*. Association of Research Libraries http://publications.arl.org/Supporting-Digital-Scholarship-SPEC-Kit-350.

Williams, B. (2010). When will we Work Out Loud? Soon! *TheBrycesWrite* https://thebryceswrite.com/2010/11/29/when-will-we-work-out-loud-soon/ (accessed 31 March 2017).

The Humanities and History Team at Columbia University. (2013) "The Developing Librarian Project". dh+lib. 1 June 2013. http://acrl.ala.org/dh/2013/07/01/the-developing-librarian-project/ (accessed 31 March 2017b).

The Programming Historian (2012) *The Programming Historian.* http://programminghistorian.org/ (accessed 31 March 2017e).

Wilms, L. (2017). Zotero | Groups > LIBER Digital Humanities working group https://www.zotero.org/groups/liber_digital_humanities_working_group (accessed 31 March 2017f).

Your Own Personal Matrix: Generationally Mediated Realities and Digital Cultures of Choice

Adrian L. Cook
adrian.cook@tccd.edu
Tarrant Country College, United States of America

The 1999 film may have been prophetic. The Matrix has us. But our children may have the Matrix. . .under their control. And it can be personalized.

This project, now at its seed level, merges performance studies, critical media studies and a reading of Jean Baudrillard's idea of simulations and simulacrum to examine the dramatization of our social reality as it occurs through each evolution of broadcast media in the twentieth century. Since the emergence of broadcast radio, each American generation has interpreted and aesthetically crafted a consensus reality that, with the onset of cable TV and the internet, has become somewhat of a simulation. This project claims that Generation Z (born after 2001) will inherit a social reality that is so highly filtered and aesthetically shaped as to be more of an abstraction of objective truth than narrative drama or even video games. Furthermore, the additional mediation of algorithms designed initially for marketing purposes will provide for Generation Z "realities of choice."

This investigation began as I observed my young daughter asserting her perceived right to choose her digitally delivered entertainment from a seemingly endless menu. The Gen Xers (my generation) are in charge of producing these entertainments, while Millennials form the bulk of the writers. This is strange alchemy, as Gen X produces both documentary and narrative drama that critiques the mythical post-war reality created by Baby Boomers, Millennials internalize and interpret that critique, producing infinite variations on themes by virtue of the sheer number of digital outlets and accessible, user friendly storytelling tools. The conceptual model of generational mediation to be presented, shows the gradual movement away from a relatively objective (or let us say homogenized consensus) social reality to and a multi-cyber-verse of highly mediated realities. The irony is that, while the beauty of this techno-narrative state allows an exponentially diversifying population space and accessibility for disseminating their narrative and to communicate their cultural reality; the option to personalize our input, in tandem with our twenty-first century culture of convenience, means that we may never see and experience nor empathize with the lives of "Others."

This work marks a continuation of a dissertation on performing myth in virtual realities and the foundational level of a "Generational Reality Theory," an examination of meaning making, knowledge acquisition and generational ways of being as codified cultural expressions. In other words, because of our emerging ability to perform ourselves and shape our perception of reality wholesale via media spaces, and the likely development of digital (and other virtual) platforms beyond our current imagination, each subsequent generation is living, quite literally, in a different reality to the extent that cross-generational communication – and for professors, andragogy – is equivalent to cross-cultural communication.

Bibliography

Baudrillard, J. (1994). *Simularca and Simulation*. 14th ed. Trans. Glaser, S. Ann Arbor: University of Michigan Press.

Active Archives: New Modes of Interaction with Online Archives

Josh Cowls
cowls@mit.edu
Massachusetts Institute of Technology
United States of America

Evan Higgins
elh@mit.edu
Massachusetts Institute of Technology
United States of America

Kurt Fendt
fendt@mit.edu
Massachusetts Institute of Technology
United States of America

This poster presentation will introduce our Active Archives Initiative, through which we aim to develop new, dynamic ways of representing, researching, and learning about the past through digital archives. From a traditional library studies perspective, most online archives are mere digital repositories, lacking the institutional and hierarchical nature of a traditional archive. Driven by a growing number of available online resources, digital technologies and tools that foster social interaction and collaboration, and open metadata, web standards and APIs, the notion of an archive continues to shift and expand.

Yet rethinking the definition of an archive is not merely a result of the digitization of existing cultural knowledge or the advent of born-digital repositories. Walter Benjamin, for example, thought of his Arcades Project — a dynamically re-organizable archive of quotes, sources, and references, developed long before the advent of modern digital technology — as something that could not leverage its full potential by being bound into a book. More recently, Wolfgang Ernst in *Digital Memory and the Archive* (2013, 81) challenges the notion of a closed archive by pointing to the fact that "new archives are successively generated according to current needs ... the object-oriented archive thus takes shape cumulatively, entailing a shift from read-only paradigms to a generative, participative form of archival reading." Laermans and Gielen (2007), moreover, observe " a quasi-bifurcation in the ways the digital archive is observed and evaluated by traditional archivists and other archive specialists", relating to the "differences between old and new media." As these examples serve to show, not merely archives in practice, but the *idea* of an archive in theory, is constantly being redefined depending on shifting needs, audiences, and available content.

Ernst's notion of a generative, participative mode of archival reading is central to the approach we introduce here. The core purpose of our Active Archives initiative is to empower users to engage in 'story-making', by discovering, interpreting and re-organizing archived materials to construct new representations of the past. In some cases, indeed, users might have their own material to contribute, which we intend to also enable. This we propose drawing together data and documents from multiple sources to enable multiple groups of users to reassemble these materials, and in so doing, challenge existing narratives and reimagine others. In many cases, the materials being used are only available in photocopy quality at best: the archive we envisage therefore allows the opportunity for the reinterpretation of valuable but incomplete resources. Our goal is to make these archives easy to understand and enjoyable to use, and to design them with a wide range of prospective users in mind, from professional scholars to high school students.

We have selected two existing projects as prototypes for this initiative. The US-Iran Relations project, a collaboration with MIT's Center for International Studies and the partners in the US and abroad, collects and digitizes governmental documents and testimonies by policymakers relating to the diplomatic relationship between the United States and Iran during the past several decades. Our second project, the Blacks in American Medicine archive, combines tens of thousands of biographical records of African American physicians with countless primary documents associated with these practitioners to create a comprehensive archive of all black physicians within the United States from 1860 to 1980. These projects leverage interactive tools such as dynamic timelines combined with customizable faceted browsers and innovative collection and story-making tools, allowing previously untold narratives and unseen connections to emerge. This will enable, as Ernst describes, "the digital archive shift to regeneration, (co-)produced by online users for their own needs" (2013, 95) while simultaneously becoming a source of dynamic media memory.

As the nature of digital scholarship changes, we need to enable archives to change as well. As we continue to develop our Active Archives Initiative, we're putting particular emphasis on the idea that our archival content is not only accessible but, more importantly, that it is intuitive and useful to diverse audiences. To that end, we hope to create fully realized, open and collaborative spaces that foster not only the proliferation of untold stories but also the ability to interact with these slices of the past in new and innovative ways. Presenting our work at DH2017 will also allow us to gather useful feedback from other practitioners in the field of Digital Humanities looking to similarly shake up the staid notion of an archive. As such, our poster presentation will offer both technical details, such as screenshots of the prototype in action, as well as clear description of the functionality and responses to our user testing.

Bibliography

Ernst, W. (2013). Digital memory and the archive. J. Parikka (Ed.). University of Minnesota Press.

Laermans, R., & Gielen, P. (2007). The archive of the digital anarchive. *Image and Narrative*, 17(April).

DHConnections: Examining the Growth of the Digital Humanities Summer Institute Community

Cole Crawford
cole.crawford@oregonstate.edu
Oregon State University

The Digital Humanities Summer Institute, or DHSI, provides "an ideal environment for discussing and learning about new computing technologies and how they are influencing teaching, research, dissemination, creation, and preservation in different disciplines, via a community-based approach" (DHSI). What began in 2001 as a small event at Malaspina University-College is now an annual 2-week affair at the University of Victoria which offers dozens of sessions and attracts a global audience. The growth of DHSI over the past 16 years mirrors the larger development of digital humanities as an academic (inter)discipline that has shifted from a niche endeavor for computational linguists and technology early-adopters to both a mainstream methodological approach and a distributed community of practitioners.

DHConnections examines DHSI attendance data. This website functions both as a tool for research about DHSI (and thus the history of disciplinary training in the digital humanities), and as a platform for DHSI alumni to connect with other researchers with similar interests, or to reconnect with contacts from previous DHSI sessions. In this way, *DHConnections* intersects with the "Collaborators" component of centerNet's excellent *DH Commons* website, but is more explicitly focused on the distributed community of DHSI alumni.

The DHSI organizational team maintains an archive of past DHSI sessions and participants. With the permission of DHSI and the Electronic Textual Cultures Laboratory at University of Victoria, I scraped this collection and cleaned it with OpenRefine to remove typographical errors, standardize attendee and organization names, and validate the data. There is no participant data (only instructors) for 2001-2003, but since 2004 there is an accurate list matching every attendee to the session he or she attended. This information is further broken out by role – student, instructor, speaker, or staff, with numerous specific subcategories of each. From 2006 onward, attendee institutional and organizational affiliations are also included. I also built a controlled vocabulary of DH topics and manually added topics to each session based on the session title and abstract when available. Together, these components form a temporal, topical, spatial, and biographical dataset which captures the attendance data for 2,678 individuals who collectively attended 254 sessions across 4,932 instances.

DHConnections allows users to access and interpret this dataset. Researchers can access the raw data via a JSON endpoint, but *DHConnections* also features numerous interactive interfaces which provide intuitive ways to understand the growth of DHSI through what Joanna Drucker calls "visual epistemologies … ways of knowing that are presented and processed visually" (2014, 8). These include a searchable table of all participants; charts of the growth of DHSI over time by country, and by total institutional attendance; a facetable map of participants' institutional affiliations; a graph which examines the popularity of DHSI session topics and when they were introduced (figure 1); and a searchable network graph of participants, linked by session attendance, scaled by frequency, and color-coded by participant role. These and other visualizations provide information about the growth of an international DH and DHSI constituency, identify major contributors and organizations within the field, and assist DH researchers interested in finding contacts.

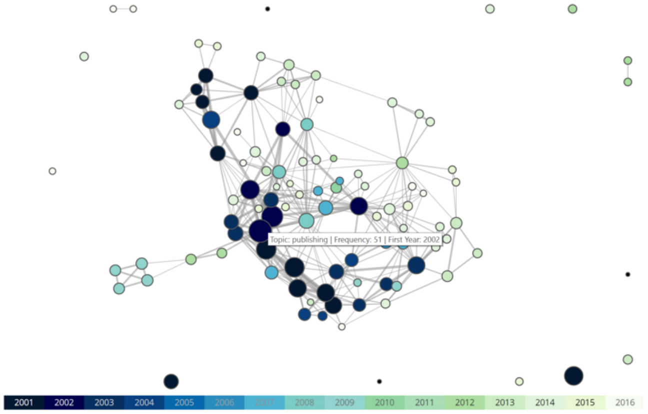


Figure 1: Topics assigned to DHSI sessions, colored by year introduced and sized by frequency

DHConnections also helps users connect with fellow DHSI alumni and find potential collaborators. For example, users interested in organizing a maker fair could query the database for DHSI alumni within 200 miles who attended a session between 2010 and 2016 and are interested in the topics of "3D printing," "augmented / virtual reality," "physical computing," or "maker culture and praxis." While "interests" are currently established by proxy via a link to attended session topics, the accuracy of such a search will increase over time because users can claim their profile on the site via an opt-in process.

Once a user claims his or her profile, he or she can edit their institutional affiliations; research interests; projects, papers, or personal websites; and/or contact information (Twitter / email). Users are also able to quickly opt-out of *DHConnections* and anonymize their attendance. *DHConnections* is designed for eventual expansion based on the availability of additional data sources such as lists of members or participants in THATCamps, HASTAC, HILT, European Summer University, DH@Madrid, DHOxSS, and conferences such as ADHO.

This poster presentation will allow ADHO attendees to test *DHConnections*. As the creator and developer of the project, I will be able to answer any questions about *DHConnections* and the underlying dataset, and look forward to talking to users and gathering feedback to help improve the platform.

Bibliography

DHSI. (2017). Digital Humanities Summer Institute Homepage. http://dhsi.org/ (accessed March 30 2017).

Drucker, J. (2014). *Graphesis: Visual Forms of Knowledge Production.* Cambridge, MA: Harvard University Press.

Beyond the Historic Facade: Skyscrapers, Scapegoats, and the Digital Reclamation of Toronto's Queer Streetscapes

Constance Crompton
constance.crompton@ubc.ca
University of British Columbia, Canada

Michelle Schwartz
michelle.schwartz@ryerson.ca
Ryerson University, Canada

Gentrification takes many forms, ranging from whole scale bulldozing and forced relocations, to skyrocketing rents and dominion of multinational corporations over entire neighbourhoods. In Toronto, gentrification is facadist: the hollowed brick skins of Victorian storefronts are preserved in orderly rows, with great glass office buildings or condos sprouting up through their once-solid roofs and porticos. Luxury boutiques and art galleries take over the spaces once held by convenience stores and leave the original, rusting signage above the door, urging pedestrians to drink Coke. A local diner closes one night and reopens the next as a cocktail lounge - the interior and exterior remain untouched. The same 1960s banquets that once served $5 burgers to an impoverished and marginalized local population now serve burgers that cost $17, even though the name advertised on the now-vintage neon sign outside remains the same.

Most city dwellers are familiar with these external signs of gentrification. Gentrification, however, does not just shape the streetscape, it shapes the outlook of city dwellers in a way that obliterates urban history, especially the history of marginalized communities, even though it is these diverse populations of people of colour, immigrants, queers, and the working poor that gave cities the dynamic qualities that made them initially so attractive to new residents. In Toronto, the building boom of the last decade has left only the facade of much of the city's diverse history. What is behind the facade - luxury condos and chain stores - represents a completely different model for a city, one that prioritizes neoliberalism, conservative values, and the homogenization of thought.

In the traditional narrative of gentrification, gentrification begins with the influx of white gay and artistic communities, the so-called "creative class," into economically disadvantaged areas. These "shock troops" of gentrification clear the way for yuppies with their preference for boutique coffee shops and organic vegetables. Rents rise and the original low-income or minority population of the neighbourhood is displaced, pushed further and further to the margins as the initial hot spot of gentrification rapidly expands. According to Sarah Schulman, Distinguished Professor of the Humanities at the College of Staten Island, this traditional narrative does not get to the heart of gentrification. The gay community has become a convenient straw man for an outcome that is shaped by much more powerful forces in municipal government, and the changing tenor of city life is due not to the introduction of artists spaces, but to the "gentrification of the mind" or the suburbanization of our understanding of city space.

In her book, The Gentrification of the Mind, Schulman quotes the artist Penny Arcade's piece "New York Values": "There is a gentrification that happens to buildings and neighborhoods and there is a gentrification that happens to ideas." Gentrification, according to Schulman's theory, is caused not simply by the influx of new people, but when those people come to cities "not to join in or to learn and evolve, but to homogenize," bringing "the values of the gated community and a willingness to trade freedom for security" (30). Gentrification "replaces most people's experiences with the perceptions of the privileged and calls that reality" and as a result "gentrified happiness is often available to us in return for collusion with injustice" (161, 166).

In the remaking of Toronto as a city of facades, are we not only erasing our history, but permanently altering the idea of what our city is, who it is for, and what purpose it serves?

And is there a way the digital humanities scholars in particular can fight these forces to a halt? Drawing on Schulman's argument and using the Lesbian and Gay Liberation in Canada project's spatial model of the Toronto gay scene as a case study, the short paper proposed here will argue that the freely accessible publication of digital models of historical communities can be used to push back against the gentrification of the mind. The presentation that will accompany the paper will mark the launch of the Lesbian and Gay Liberation in Canada map, built from the 17,000+ person, place, and event entity records in the LGLC database.

The Lesbian and Gay Liberation in Canada project reconfigures Donald McLeod’s remarkable two-volume work, Lesbian and Gay Liberation In Canada: A Selected Annotated Chronology, 1964-1981 as a TEI-encoded resource and graph database. The prototype database, available at http://lglc.ca consists of event, publication, person, and place records spanning from the founding of the first homophile associations in Canada through to the start of the AIDS crisis. The LGLC project extends McLeod’s codex form in order to data mine and represent queer history spatially and temporally. The project not only makes a much-neglected part of Canadian history available for mainstream scholarly use, it also provides a foundation for modeling identity and representing time and space in TEI.

There is much to model spatially and temporally. In the 1960s, the Toronto’s gay scene was located on Yonge Street and by the 1970s the bars had migrated north along Yonge from Queen Street towards Wellesley Street. Two of the most famous bars on this strip were the Parkside at 530 Yonge Street and the St. Charles Tavern at 488 Yonge Street, both owned by Norman Bolter. The neighbourhood surrounding the former sites of the Parkside and the St. Charles is currently a site of massive development. Gone are the seedy bars and massage parlours. Large glass condos are rising from bulldozed lots and from the tops of heritage buildings.

The building that housed the St. Charles Tavern still stands above Yonge Street at Wellesley. It is slated to become a condo in 2017, the iconic clock tower that served for decades as a meeting spot for the gay community will be preserved at the base of a proposed 153 metre glass spire. The incorporation of this piece of gay history into a shining skyscraper only furthers serves the ideology that scapegoats the gay community as the root cause of gentrification, making the million dollar condos imagined by a developer seem like an inevitability, rather than just one potential future amongst many. What is lost in this new instance of facadism is the history of the St. Charles Tavern, a violent history neither the city nor developers seem keen to acknowledge or commemorate.

For example, tracing the events outside the St. Charles Tavern on Halloween from 1968 to 1981 reveals Toronto's violently homophobic history, and the inadequate policy and police response to that violence. In the 1960s it was illegal for men to wear clothing of the opposite sex in Toronto, a rule that was relaxed at Halloween. Every Halloween St. Charles Tavern hosted a drag show, and every year a homophobic mob congregated outside the bar to throw eggs and rotten vegetables as the queens arrived. In 1968, police found gasoline bombs behind the tavern, and yet in the following years police still permitted crowds to hurl insults shout slogans such as "kill the queers", and throw debris at the patrons. It took a decade for the police to start arresting people in the mob; in the meantime a volunteer gay defense patrol, Operation Jack O'Lantern, escorted patrons through the neighbourhood and in through the back door of the tavern. Between 1975 and 1978, 14 patrons of Toronto's gay bar scene were murdered. Eight of those murders went unsolved amid accusations of police homophobia, which were only exacerbated by homophobic articles published in the Toronto Police Association Journal. Publishing this history online, not only in text, but in map form, will make it accessible to an audience that will include not just scholars, but the citizens of Toronto. By developing accompanying pedagogical tools to guide users through the map, we hope to engage local residents, as well as the thousands of students at nearby universities and schools, in working to reclaim that lost history, and undoing some of the effects of the gentrification of the mind.

The Bodleian TEI Catalogue Consolidation Project

James C. Cummings
james.cummings@it.ox.ac.uk
University of Oxford, United Kingdom

Andrew Hankinson
andrew.hankinson@bodleian.ox.ac.uk
University of Oxford, United Kingdom

The Bodleian’s TEI Catalogue Consolidation project was a joint Bodleian / IT Services project at the University of Oxford, designed to implement a single consolidated and sustainable infrastructure for delivering TEI manuscript description catalogues. The project incorporated user research, scoping, technical design, implementation, user testing, and subsequent iteration. The aims of the project were to determine a suitable technical architecture for the storage and indexing of TEI manuscript descriptions across multiple collections; engage in user testing of the existing TEI-based catalogues at the Bodleian and decide on the functional improvements required for the front end interface(s); build or implement the ‘back-end’ technical architecture as scoped; build a new, user-friendly interface for searching and browsing TEI; and design and implement a sustainability plan around training, communications and standards.

This poster will concentrate on the technical framework, joint schema design, and legacy data migration of existing catalogues and where this fits in the overall technical systems of the Bodleian Library’s consolidated infrastructure for TEI manuscript description catalogues.

The Bodleian Libraries holds well over 20,000 manuscripts and serves readers from all academic divisions of the University of Oxford as well as thousands of external readers from around the world. More than half of the manuscript collections in the Bodleian are effectively hidden from scholars because they are not described online, or because the existing descriptions are not well indexed or even available to users. To try to solve this problem, the Bodleian spent five years creating TEI-based catalogues for descriptions of its manuscripts. However, the result was eight separate TEI catalogues, the largest of which (FIHRIST) is the union catalogue for the description of Islamic manuscripts in the UK. Seven other catalogues (Hebrew, Armenian, Georgian, Tibetan, Sanskrit, Shan, and Genizah) represent portions of the Bodleian Libraries Oriental collections. The most recent addition is the TEI catalogue for western medieval manuscripts, funded by *The Tolkien Trust*. Prior to the consolidation project, small improvements were made to each subsequent catalogue, but funding was not available to rollout those improvements retrospectively, or to consolidate the code base. TEI is rapidly becoming the

de-facto language for encoding descriptions of manuscripts collections (in the Bodleian and internationally) and this project was an opportunity to build on its use in a concrete way, by providing a clear set of software, guidelines, and training capacities for the large number of staff engaged with describing manuscripts. Consolidating this infrastructure has significantly lowered the operating costs of support and feed directly in the University's strategy for enhancing access to collections.

As part of the project, the existing TEI manuscript description schemas were consolidated into a single TEI customization. By using TEI, the catalogue descriptions were made more easily expandable to include full transcriptions of manuscripts, leading the way towards the mark-up and extraction of important biographical, geographical, and scientific data. In addition the existing catalogue descriptions were migrated to this single TEI customization. As well as showcasing the project as a whole, the poster will focus on the consolidation process including the key aims and objectives of the project, which were:

- To decide on a technical architecture for the storage and indexing of TEI-XML manuscript descriptions across multiple collections.
- Engage in user testing of the existing TEI-based catalogues at the Bodleian to decide the functional improvements required for the front end interface(s).
- Develop a technical requirements document with the necessary details to implement the recommendations during the subsequent buildproject.
- Build the 'back-end' technical architecture as scoped.
- Build a new user-friendly public interface for searching and browsing the TEI.
- Migrate the existing TEI applications into the new solution.
- Design and implement a sustainability plan around training, communications and standards.

The poster will explain the project, the process used for schema consolidation, legacy data migration, and the overall technical infrastructure.

“My Name is Lizzie Bennet”: Reading, Participation, and Jane Austen Across Media Platforms

Meredith Dabek
meredith.dabek@nuim.ie
Maynooth University, Ireland

The Lizzie Bennet Diaries is a digital update of Jane Austen's novel, *Pride and Prejudice*, in which Austen's narrative is reimagined for the twenty-first century through its distribution across multiple media platforms. *The Lizzie Bennet Diaries* (hereafter referred to as LBD) is centred on a series of YouTube video diaries by Lizzie Bennet and includes four complementary YouTube channels, thirteen interconnected Twitter feeds, Tumblr posts, Facebook profiles, and numerous social media interactions and 'conversations' between the narrative's characters and its readers.

This poster expands upon existing research conducted as part of a larger, ongoing PhD research project. It will present a visual overview of LBD and the specific modes of participation available to readers during the narrative's initial release in 2012 and 2013 (an important distinction, since the narrative was originally released as a serial story). Drawing from a sample of LBD reader comments on YouTube and Twitter (using MaxQDA to qualitatively code the comments), the poster will explore how readers participated in the LBD narrative and how their participation may have influenced or affected their reading habits. In addition, Aarseth's cybertext theory and McGann's radial reading theory will provide a foundation for discussing LBD's participatory elements in a theoretical context, with an emphasis on discussing how digital media have invited us to revisit and rework those theories.

According to Aarseth, cybertexts such as LBD invite and encourage readers to make deliberate and intentional choices as they navigate through the text, its multiple entry points, and its various narrative paths (1997). As a result, readers actively participate in shaping their individual reading experience. Thus, LBD reader must decide which elements of the narrative to consume, and in which order: does she choose to follow Lizzie on Twitter, but not Darcy? Does she watch Lydia's YouTube videos, or reblog Jane's fashion posts on Tumblr? Each choice leads the reader in a slightly different direction in the narrative, and provides additional information that offers context and meaning to the core series of YouTube videos. The choices made by readers are also indicative of McGann's theory of radial reading, where readers seek out additional information not immediately available in the core text (McGann 1991). The various modes of participation available to LBD readers, as well as the degree of interactivity implied by cybertext theory and radial reading, gives readers an avenue for active engagement with the narrative and helps reinforce the belief that their contributions to the narrative matter. This belief, what Professor Stephen Coleman calls "*the feeling of being counted*, or the affective character of an experience that renders it fulfilling for individuals" (2013), serves to strengthen a reader's engagement with the text.

Using Aarseth and McGann's theories as a foundation, the poster will consider how LBD's modes of participation might strengthen a reader's engagement with and immersion in the text, and whether immersive digital narratives like LBD can encourage engagement with traditional close reading techniques by prompting individuals to read (or reread) Austen's original *Pride and Prejudice* novel. As Frank Rose, Henry Jenkins and others have pointed out, the Internet has changed readers' expectations from stories and narratives, to the point that many readers of digital narratives now expect some level of participation or interactivity. This poster (and the larger research project) is focused on exploring those participatory elements and their connection to overall digital reading practices.

Bibliography

Aarseth, E. (1997). *Cybertext*. Baltimore: Johns Hopkins University Press.

Coleman, S. (2013). *How voters feel*. Cambridge: Cambridge University Press.

Dowling, D. (2014). "Escaping the Shallows: Deep Reading's Revival in the Digital Age." *Digital Humanities Quarterly*, [online] 8(2). Available at: http://www.digitalhumanities.org/dhq/vol/8/2/000180/000180.html [Accessed: 27 Oct. 2016].

Jenkins, H. (1992). *Textual Poachers*. New York: Routledge.

McGann, J. (1991). "How to Read a Book" in *The Textual Condition*. Princeton: Princeton University Press.

Pemberley Digital. (2016). *The Lizzie Bennet Diaries*. [online] Available at: http://www.pemberleydigital.com/the-lizzie-bennet-diaries/ [Accessed: 27 Oct. 2016].

Towards an Integrated Set of Annotations for Folktales

Thierry Declerck
declerck@dfki.de
German Research Center for Artificial Intelligence University of Vienna, Austria

Introduction

In this poster paper we describe very briefly different layers of annotation for folktales we have been working in the past and which are in the process of being integrated in one set of annotations, which is mediated by a formal representation of the annotation elements in an ontological framework. We list in this short text the various modules of this integrated annotation scheme.

A first approach to the annotation of folktales was done in the context of cooperation between the European CLARIN project and the Dutch Amicus (*Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts,* sponsored by a grant from the Netherlands Organization for Scientific Research, NWO Humanities, as part of the Internationalization in the Humanities programme, from 2009-2012). In this context, we developed an extended annotation scheme for the annotation of folktales with Proppian functions. The scheme includes textual properties, temporal structures, characters, dialog structures, and the Proppian functions (see Declerck et al, 2011 and Scheidel and Declerck, 2010). This scheme was later used for supporting a first information extraction system applied to tales (see Declerck and Scheidel, 2010), and comparing the results of this extraction with manually annotated tales.

Building on this work, an automated linguistic analysis of tales was developed. The goal was not only to detect characters of the tales, but also to provide for a co-reference analysis such that the actions in which the characters are involved can be fully specified, and thus helping for an automated detection of Proppian functions, together with the involved personages. Results of the analysis are stored in a database, which has been further developed onto an ontological framework: Adding thus not only an annotation layer but also a formal representation level (see Koleva et al, 2012, and Declerck et al, 2012). The ontological representation allows also to apply generalisations for the specification of the characters (human vs animals, or supra natural etc.). The system was also able to operate reference resolution of the kind: "daughter" can also be a "sister", etc.

The move to the use of an ontological framework turned out to be very useful, since further work on distinct elements of a tales could be easily integrated. So for example, the work described in Eisenrecih et al (2014) considered the detection of sentiments expressed by the characters of the tales. Such sentiments ("joy", "happiness", "sadness" etc.) could be added in a straightforward manner to instances of characters at the time span in which they occur in the tales. In fact, the work in Eisenrecih et al (2014) mainly addresses the issue of adding Text to Speech (TTS) functionality to the automatic analysis of the text. The TTS system accesses the instances of the characters in the populated ontology and can retrieve the information on sentiment encoded there and correspondingly model the voice output of the various characters (an example using the tale "Frog Prince" can be heard, or preferably downloaded, online).

Very recently, we started also looking at other metadata to be used for annotating folktales, and to see how to integrate those with the Proppian functions. We looked for this at the well-known classification systems of Stith Thompson (1977), Antti Aarne (1961) and Hans-Jörg Uther (2004) and we are starting to integrate those models in our ontology. The resulting ontology from the Thompson Motif Index has been presented in Kostova and Declerck (2016). Additionally we linked the detected characters to WordNet, investigating if this can help for the disambiguation of characters (Declerck et al, 2016).

As the most recent work we dedicated to the folktale topic dealt with with the implementation of running systems, less attention was given into the extension of the annotation scheme, work which is currently underway and which we present as a poster at DH 2017.

Bibliography

Declerck, T., Scheidel, A., Lendvai, P. (2011) Proppian Content Descriptors in an Integrated Annotation Schema for Fairy Tales. *Language Technology for Cultural Heritage. Selected Papers from the LaTeCH Workshop Series,Theory and Applications of Natural Language Processing, Pages 155-169, Springer, Heidelberg, 2011*

Declerck, T., Scheidel, A. (2010) An Information Extraction Approach to the Semantic Annotation of Folktales. In: Sándor Darányi, Piroska Lendvai (eds.): *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, Vienna, Austria, 10/2010

Scheidel, A., Declerck, T. (2010) APftML - Augmented Proppian fairy tale Markup Language. In: Sándor Darányi, Piroska Lendvai (eds.): *First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts,* Vienna, Austria, 10/2010

Declerck, T., Scheidel, A., Lendvai, P. (2010) Proppian Content Descriptors in an Augmented Annotation Schema for Fairy Tales. In: Caroline Sporleder, Kalliopi Zervanou (eds.): *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Lisbon, Portugal, IOS Press, European Coordinating Committee for Artificial Intelligence -- ECCAI, 8/2010*

Koleva, N., Declerck, T., Krieger, H-U. (2012). An Ontology-Based Iterative Text Processing Strategy for Detecting and Recognizing Characters in Folktales**.** In: Jan Christoph Meister (ed.): *Digital Humanities 2012 Conference Abstracts, Pages 467-470, Hamburg.*

Declerck, T., Koleva, N., Krieger, H.-U. (2012). Ontology-Based Incremental Annotation of Characters in Folktales. in: Kalliopi Zervanou, Antal van den Bosch, (eds.): *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH 2012) , Pages 30-35, Avignon, France, 4/2012

Eisenreich, C., Ott, J., Süßdorf, T., Willms, C., Declerck, T. (2014). From Tale to Speech: Ontology-based Emotion and Dialogue Annotation of Fairy Tales with a TTS Output *Proceedings of ISWC 2014,* Riva del Garda, Italy, Springer.

Declerck, T. (2015). **Annotationen für die automatisierte Verarbeitüng von Märchen. In:** *Book of Abstracts, DHD 2015***,** Graz, Austria.

Thompson, S. (1977). *The Folktale.* Berkeley: University of California Press

Aarne, A (1961). *The Types of the Folktale: A Classification and Bibliography.* The Finnish Academy of Science and Letters, Helsinki.

Uther, H-J. (2004). The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson. FF Communications no. 284–286. Helsinki: Suomalainen Tiedeakatemia.

Declerck, T., Klement, T., Kostova, A. (2016). Towards a WordNet based Classification of Actors in Folktales. In: Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, Piek Vossen (eds.): *Proceedings of the Eighth Global WordNet Conference*, Bucharest, Romania, GWA, Global WordNet Association, 1/2016

Kostova, A., Declerck, T. (2016). Ontologisierung vom Thompson Motif's Index. In: *Book of Abstracts of DHD2016*, Leipzig, Germany

Creating a Policy Framework for Analytic Access to In-Copyright Works for Non-Consumptive Research

Eleanor Dickson
dicksone@illinois.edu
University of Illinois, United States of America

Daniel G. Tracy
dtracy@illinois.edu
University of Illinois, United States of America

Sandra McIntrye
mcintsan@hathitrust.org
HathiTrust Operation, United States of America

Bobby Glushko
rglushko@uwo.ca
University of Western Ontario, Canada

Robert H. McDonald
rhmcdona@indiana.edu
Indiana University, United States of America

Brandon Butler
bcb4y@eservices.virginia.edu
University of Virginia, United States of America

J. Stephen Downie
jdownie@illinois.edu
University of Illinois, United States of America

Introduction

We report on the work of a recent HathiTrust Research Center (HTRC) task force charged to draft an actionable, definitional Non-Consumptive Use Research Policy. As the research division of HathiTrust, the HTRC facilitates computational text analysis of materials in the HathiTrust Digital Library (HTDL) by adhering to a non-consumptive research paradigm. As the HTRC has integrated the text of the full HTDL corpus into its datastore, it has become increasingly important to clarify and codify the Center's policy for non-consumptive research. The task force, which consisted of copyright and scholarly communications librarians and representatives from HathiTrust operations and the HTRC, recommended a policy that clarifies acceptable researcher behavior and allowable exports from the HTRC Data Capsule (Plale, et al., 2015). This poster describes the task force's work to establish a Non-Consumptive Use Research Policy for the HTRC that aims to achieve the same goals as copyright itself: to promote progress in the discovery and spread of knowledge, without harming the commercial interests of authors, publishers, and other stakeholders.

Background

While the concept of non-consumptive research has seeded the mission of the HTRC, the Non-Consumptive Use Research Policy task force sought to translate conceptual definitions of the term into practicable policy (Bhattacharyya, et al., 2015). When first used in 2010, the term was defined as "research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book" (Amended Settlement Agreement: Authors Guild, Inc., et al., v Google Inc., 2009). Since then, legal scholar Matthew Sag and literary scholar Matthew Jockers have offered their own definitions and assessments, tending to favor instead the term *non-expressive use* (Sag, 2012; Jockers, 2013). Several recent court decisions pointed the task force toward the current legal understanding of non-consumptive research specifically and current interpretations of fair use broadly (Authors Guild v HathiTrust, 2014; Cambridge University Press v Becker, 2016; Fox News Network v TV Eyes, 2014). Additionally, the task force looked to existing access models for restricted data (ICPSR) as well as professional guidelines for non-consumptive research (Association of Research Libraries, 2012; Cox, 2015).

Policy Highlights

The group first created a framework drawing on fair use that, when paired with the HTRC technical infrastructure, would clarify non-consumptive access to the HTDL. This framework accounted for several considerations and safeguards:

- Mechanical data mining differs from researcher-driven computational text analysis, which requires interplay between scholar and text.
- Current case law suggests that it needs to be sufficiently difficult, but not strictly impossible, to reconstruct the expressive work (Authors Guild v Google Books, 2015).
- Users must agree that they will not treat HTRC tools as a reading application, and the tools should periodically remind them of this limitation.
- The HTRC must continue to block through technological measures and human review the export of protected textual data from the secure system.

The task force then drafted the HTRC Non-Consumptive Use Research Policy (HathiTrust, 2017). It defines non-consumptive research as "Research in which computational analysis is performed on one or more volumes or textual objects in the HTDL, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted work to understand the expressive content presented within that work." Of key importance is the notion of substantial portion, which, according to the policy, is a portion of the work sufficient in quality or quantity to provide a substitute for access to the expressive content of the original text. The policy outlines acceptable in-capsule uses of corpus text that are limited to those which would facilitate scholarly text analysis, including checking results to refine algorithms. In addition to enumerating non-consumptive research practices—for example text extraction, textual analysis, and automated translation—the policy provides sample results that further model approved uses. These results, which may be exported from the HTRC Data Capsule, include non-binary, human-readable statistical summaries, derived results, keywords-in-context, and concordances that are not sufficient to reconstruct a substantial portion of the text.

Conclusions

The task force tailored the policy to address current infrastructure within the HTRC, both technical and human, as opposed to accounting for prospective updates to interface and design. As such, the policy is an iterable, living document that must be revisited as HTRC systems are further developed. Such technical developments, such as the HTRC's exploration of machine-aided results verification to augment the current human-review system, will improve the scalability of the HTRC Data Capsule and may require updates to the policy. As more researchers interact with the HTRC Data Capsule, their use cases may prompt additional refinement of the policy, especially in the exemplar results it provides. The process followed in developing the policy, as well as the guidelines themselves, may be useful in other text mining research environments. They encourage an interpretation of non-consumptive research that values scholarship and intellectual progress, while still balancing the restrictions imposed by copyright law.

Acknowledgements

Task force members included: Aaron Elkiss, Brandon Butler, Bobby Glushko, Daniel G. Tracy, Eleanor Dickson, Robert McDonald, and Sandra McIntyre.

Bibliography

Amended Settlement Agreement: Authors Guild, Inc., et al., v Google Inc. (2009).

Association of Research Libraries (2012). *Code of Best Practices in Fair Use for Academic and Research Libraries.* Available from: http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf (accessed 1 November 2016).

Authors Guild v Google Books (2015).

Authors Guild v HathiTrust (2014).

Bhattacharyya, S., Organisciak, P., and Downie, J. S. (2015). "A fragmentising interface to a large corpus of digitized text." *Interdisciplinary Science Reviews,* 40(1): 61-77.

Cambridge University Press v Becker (2016).

Cox, K. L. (2015). "ARL Issue Brief: Text and Data Mining and Fair Use in the United States," Available from: http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf (accessed 1 November 2016).

ICPSR (2017). "Data Enclaves," *University of Michigan*, Available from: http://www.icpsr.umich.edu/icpsrweb/content/icpsr/access/restricted/enclave.html (accessed 1 November 2016)/

Fox News Network v TV Eyes (2014).

Jockers, M. (2013). *Macroanalysis: digital methods and literary history.* Champaign: University of Illinois Press.

HathiTrust (2017). "Nonconsumptive Use Research Policy." Available from: https://www.hathitrust.org/htrc_ncup (accessed 17 March 2017).

Plale, B., Prakash, A., and McDonald, R. (2015). "The Data Capsule for Non-Consumptive Research: Final Report."

Sag, M. (2012). "Orphan Works as Grist for the Data Mill." *The Berkeley Technology Law Journal* 27: 1503-50.

Neoclassica – an open framework for research in Neoclassicism

Simon Donig
simon.donig@uni-passau.de
University of Passau, Germany

Maria Christoforaki
maria.christoforaki@uni-passau.de
University of Passau, Germany

Siegfried Handschuh
siegfried.handschuh@uni-passau.de
University of Passau, Germany

Bernhard Bermeitinger
bernhard.bermeitinger@uni-passau.de
University of Passau, Germany

In the History of Art, the shaping of aesthetic forms, the transfer of the stylistic languagem and the change of styles have always been at the heart of the discipline. The co-occurrence of aesthetic forms in multiple modi (such as architecture, applied arts, graphics and textual description) or in larger bodies of artefacts has been researched to a lesser degree. Instruments and methods from the field of the Digital Humanities are now opening up pathways for a new understanding of such issues by supporting hermeneutic understanding with quantitative means.

Such instruments are particularly valuable for researching far reaching aesthetic movements such as Neoclassicism (ca. 1760-1860). The Neoclassic movement was of almost global scale – affecting architecture and design from Sidney to New York, and from Athens to the outreach of the Russian Urals – while relating to a common reference in Classical Antiquity, therefore making it an almost ideal topic for studying processes of stylistic transformation.

Here, we present the Neoclassica research framework (The Neoclassica Project, 2017), aimed at providing historians and art historians with new instruments and methods for analysing and classifying material artefacts and aesthetic forms. Initially we are focusing on works of applied art (for instance furniture and furnishings) as well as architecture, following the preliminary hypotheses that these modi show aesthetic forms in close communication with one another due to constructional commonalities and their shared reference of the Classical. While these forms show a trove of similarities across a huge geographic range they also diverge due to local traditions, needs and limitations or political conditions supporting the use of particular vocabularies of aesthetic forms.

We consider it most promising to follow these trajectories in a fresh way, applying the instruments provided by the Neoclassica system, joining a top-down approach harnessing the domain expert knowledge and a data-driven bottom-up approach exploiting the powers of computationally processing large amounts of data.

Specifically, the top down approach comprises a formalized knowledge-representation based on a newly developed ontology for classical artefacts and the semantic annotation and curation of a specifically developed corpus of images and texts done by experts.

The Neoclassica ontology aims to establish a controlled vocabulary that will be both research oriented and multilingual, while at the same time taking into account the different shape of the represented concepts in different languages, specifically in English, French and German (Donig, Christoforaki & Handschuh, 2016). The concepts and terms used to describe artefacts and their structure are based on period sources such as such as for instance Charles Percier's and Pierre François Léonard Fontaine's or Thomas Sheraton's pattern-books or George Hepplewithe's treatise on interior decoration (Hepplewhite, 1794; Sheraton, 1802, 1803; Percier, 1812) reflecting the conceptual control over the production of artefacts.

Since January we started the development of a semantic annotation tool for image corpora. The initial corpus sources will be commercial providers like Auction Houses but we also strive to include coherent ensembles. For this we have for instance a partnership for digitizing the interiors and furnishings of the Dessau-Wörlitz UNESCO World Heritage Site. The digitization and annotation will be done in collaboration with the chair for Visual Studies and Art History of our University. We conceptualize the process of corpus annotation to be in a constant dialog with the ontology development, so that the findings of domain experts curating the data will enrich the perception for the overall domain.

The bottom-up approach, on the other hand, is data-driven in the sense that it employs Deep Learning and Distributed Semantics algorithms for knowledge extraction and classification from images and text corpora (Bermeitinger, Donig, Christoforaki, et al., 2017). We apply a Deep Learning algorithm for classifying features in digital images of classical artefacts like furniture, works of applied art and architecture. We tested the algorithm in a trial image corpus with coarse label annotation for the image as a whole (The corpus and accompanying source code can be downloaded from The Neoclassica Project, 2017). Currently we are slowly progressing to a more refined feature annotation trained on a corpus annotated by experts using the ontology.

Knowledge discovery in text corpora is divided into Named Entity and Relationship extraction mechanisms, already robust in English and rapidly evolving in the German language. This is complemented with a schema agnostic natural language query interface. These modules are part of the in-house developed open source system StarGraph (Lambda3 Project, 2017) capable of processing both structured and unstructured data.

We aspire for the Neoclassica framework to become the centre of a sustainable, open community of scholars with a multitude of disciplinary backgrounds from both the humanities and computer sciences, making it an amalgam system that combines the best of two worlds.

Notes

The project team may be contacted at neoclassica@fim.uni-passau.de .

Bibliography

Bermeitinger, B., Donig, S., Christoforaki, M., Freitas, A., Handschuh, S. (2017) "Object Classification in Images of Neoclassical Artifacts Using Deep Learning." In: DH2017 Montréal.

Donig, S., Christoforaki, M. & Handschuh, S. (2016) "Neoclassica - A Multilingual Domain Ontology. Representing Material Culture from the Era of Classicism in the Semantic Web." In: B. Bozic, G.

Hepplewhite, G. (1794) The cabinet maker and upholsterer's guide; or, Repository of designs for every article of household furniture. London, I. and J. Taylor.

Lambda3 Project (2017) Stargraph. [Online]. 2017. Lambda3. Available from: http://lambda3.org/Stargraph/ [Accessed: 3 April 2017].

Mendel-Gleason, C. Debruyne, & D. O'Sullivan (eds.) (2016) Computational History and Data-Driven Humanities. CHDDH 2016. IFIP Advances in Information and Communication Technology, vol 482. [Online]. Cham, Springer. pp. 41–53. Available from: https://link.springer.com/chapter/10.1007/978-3-319-46224-0_5 [Accessed: 31 March 2017].

Percier, C. (1812) Recueil de décorations intérieures. Paris, Didot.

Sheraton, T. (1803) The cabinet dictionary. To which is added a supplementary treatise on geometrical lines, perspective, and painting in general. London, Smith.

Sheraton, T. (1802) The Cabinet-maker and upholsterer's drawing-book in four parts. 3rd edition. London, Bensley.

The Neoclassica Project (2017) Neoclassica – A Framework for Research in Neoclassicism. [Online]. 2017. Available from: http://neoclassica.network [Accessed: 3 April 2017].

The Neoclassica Project (2017) Ressources. [Online]. 2017. Neoclassica – A Framework for Research in Neoclassicism. Available from: http://neoclassica.network/resources [Accessed: 3 April 2017].

Authorship of Dream of the Red Chamber: A Topic Modeling Approach

Keli Du
keli.du@stud-mail.uni-wuerzburg.de
University of Würzburg, Germany

Dream of the Red Chamber (DRC, 红楼梦) is one of the most famous Chinese classic novels, written by Cao Xueqin (曹雪芹) during the 18th century. The original version of DRC had 80 chapters. But in 1791, Gao E (高鹗) and Cheng Weiyuan (程伟元) claimed that they had found more manuscripts of Cao and published another edition with 120 chapters. Since then, there has been a lot of discussions regarding the number of authors of DRC. Many scholars see the last 40 chapters as a later addition. Currently Hu Shih's (胡适) (Hu, 1988) research is most widely accepted, where he argued these last 40 chapters were written by Gao E. According to some modern research approaches, the first

80 and the last 40 chapters are written by two authors. Evidence also suggest that Chapters 64 and 67 may not be written by Cao (Hu, Wang, & Wu, 2014; Tu, & Hsiang, 2013).

Using Delta (Burrows, 2002), a measure of difference between two texts, the same conclusion has been obtained (Du, 2016). The 120 chapters are written by two authors. Red texts are the first 80 chapters and green texts are the last 40 chapters. Delta also suggest that chapter 6, 10, 11 and 67 might be written by the second author (see Fig. 1, 2, 3).

Although most the results obtained are within expectations, the presence of the four chapters in the second group certainly deserves further investigation. Three hypotheses are proposed in this paper as the cause for this situation:

- This test result from the Delta method is not 100% accurate.
- These four chapters shares many names or plot related terms with the last 40 chapters.
- Stylistic difference. Compared to the other chapters, the use of some less plot related terms indicates that the second author wrote Chapter 6, 10, 11 and 67.

Topic Modeling was used to test DRC on this regard. I used the version of the DRC that Tu (2013) deemed “the closest to the earliest editions” for this study. Topic Modeling can automatically discover the contents of a large collection of documents, and is often used as an alternative method to explore the documents. A topic is a probabilistic distribution over words appearing in the corpus. The model finds groups of related words, and words that occur frequently together will be clustered in the same group. If some words tend to co-occur in the last 40 chapters of DRC, Topic Modeling would be effective in highlighting their presence. LDA (Latent Dirichlet Allocation) (Blei et al., 2003) was used to model my corpus and do my test, and MAchine Learning for LanguagE Toolkit (MALLET) was used as the topic-modeling tool.

At the preprocessing step, tokenizing and chunking were performed. Tokenizing is required to process texts written in the Chinese Language to spilt texts into words, as there are no spaces to mark word boundaries. Tools like Stanford Chinese Word Segmenter can only be used on modern Chinese texts, as the segmentation standards are not suitable for classic Chinese. Character bigrams were hence selected as the “word”. Breaking the texts up allows the relationship among words to be explored more thoroughly, hence the chapters were split, where each document for the test contains 500 bigrams. Stopwords present another issue. A test run was first performed with MALLET to acquire some topics. The results of the test run show the bigram words assigned to the topics. In these topics the correct bigram words and person names were observed. Besides, a significant amount of meaningless bigrams in the topics were also found, for example: 著一 (writings, one), 的干 (and so on, do), 云笑 (cloud, smile). Thereby a stopword list with the following was compiled: a collection of meaningless bigrams and person names and function words like 一個 (a), 這個 (this), 只得 (have to) and so on.

MALLET was run after the preprocessing stage to generate 50 topics. The topic-document distribution output was aggregated as the chapters were split into chunks at the preprocessing step. The average of the topic shares associated with each chapter were then computed and the shares were transformed into a document-topic matrix. Then the visualization of the topic proportions associated with the chapters was created (see Fig. 4). The x-axis are the topics and the y-axis are the documents. The red line divides the figure into two parts, the first 80 chapters (above the line) and the last 40 chapters (under the line). Topic #26 became the main focus from this result: the last 40 chapters are all strongly associated with it, while the first 80 chapters (except Chapter 11 and 67) are associated to this topic with much lower probability. The topic-word assignments were then computed. The top 20 words and their unnormalized weights of the topic #26 are:

> 357.0: 太太 (lady), 246.0: 回來 (back), 233.0: 來了 (coming), 218.0: 丫頭 (slave girl), 190.0: 過來 (come over), 184.0: 答應 (answer), 146.0: 只見 (seen), 125.0: 去罷 (go), 122.0: 時候 (moment), 114.0: 叫人 (call someone), 108.0: 出來 (come out), 101.0: 言語 (speech), 100.0: 告訴 (tell), 98.0: 今日 (today), 98.0: 話儿 (talk), 95.0: 看見 (see), 92.0: 回道 (answer), 91.0: 那邊 (over there), 91.0: 連忙 (promptly), 85.0: 想起 (think of)

All the words are not person names or plot-related. The result obtained from the Topic Modelling shows that Topic #26 is a “style”-Topic. Most of them can be presented in a scene: an interaction (or a conversation) between the lady and the slave girl. DRC is an observation of the Chinese society in 18th-century. The story is about the life of two large, wealthy family compounds in the capital of China. Such scenes or plot are hence very common in the whole novel.

In conclusion, Topic Modeling was used in this paper to find the specific topic, which represents the difference between the first 80 and the last 40 chapters of DRC. The test results indicate that both hypotheses, a) the Delta test result is not 100% accurate, b) The four chapters share many names or plot related terms with the last 40 chapters, are not true. The first 80 chapters (except Chapter 11 and 67) are stylistically different from the last 40 chapters. According to the results of Delta and Topic Modeling, both Chapter 11 and 67 are definitely not written by the first author. They might be written. or at least edited by the second author of DRC.

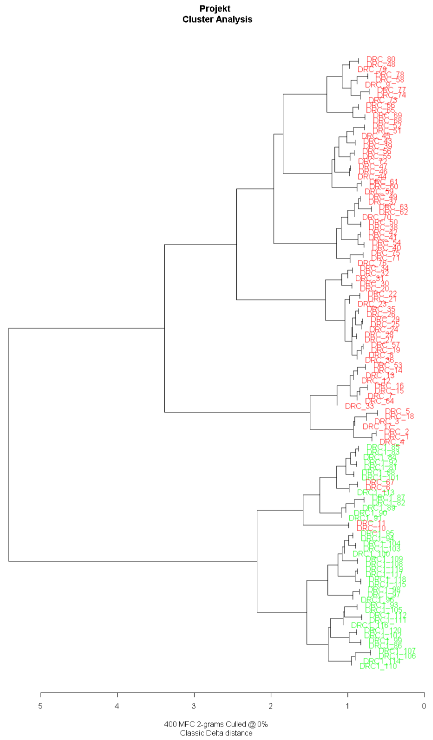


Figure 1. Delta test results of DRC, (300 MFC, 2-grams)

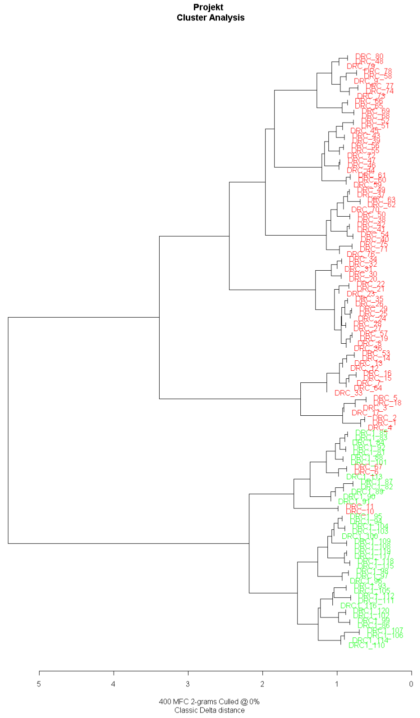


Figure 2, Delta test results of DRC, (400 MFC, 2-grams)

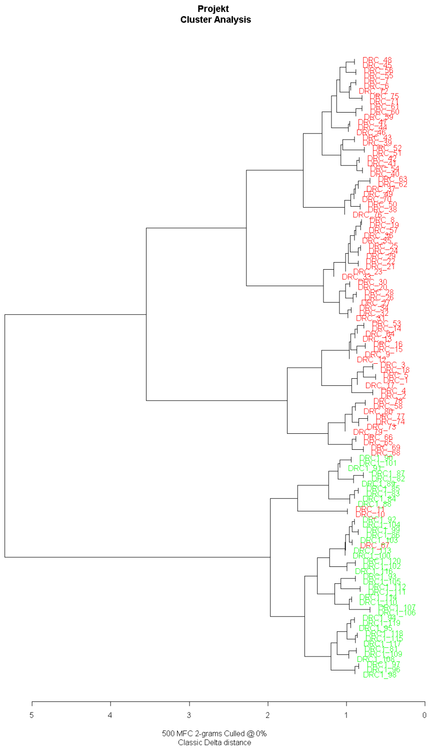


Figure 3. Delta test results of DRC, (500 MFC, 2-grams)

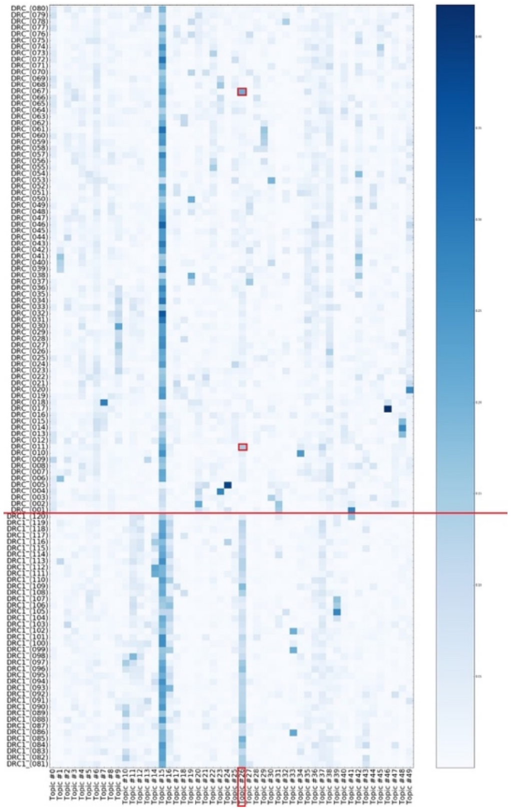


Figure 4. Topic-chapter distribution of DRC (50 topics, 120 documents)

Bibliography

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 993-1022.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. In: Literary and Linguistic Computing, 17(3), (pp. 267-287).

Du, K. (2016). Testing Delta on Chinese Texts. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 781-783.

Hu Shhi (1988). 《胡适红楼梦研究论述全编》 [Hu Shihs Analysis of Dream of Red Chamber], Shanghai Guji Chubanshe (Shanghai Classics Publishing House)

Hu, X., Wang, Y., & Wu, Q. (2014). Multiple Authors Detection: A Quantitative Analysis of Dream of the Red Chamber. Advances in Adaptive Data Analysis, 1450012.

Tu, H. C., & Hsiang, J. (2013). A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red Chamber. In: Digital Humanities 2013: Conference Abstracts (pp. 441-443)

Introducing the Open Online Newspaper Initiative

Jessica Dussault
jdussault@unl.edu
University of Nebraska, United States of America

Laura Weakly
lweakly2@unl.edu
University of Nebraska, United States of America

Karin Dalziel
kdalziel2@unl.edu
University of Nebraska, United States of America

Jeremy Echols
jechols@uoregon.edu
University of Oregon, United States of America

Karen Estlund
kme20@psu.edu
Pennsylvania State University
United States of America

Andrew Gearhart
andrew@psu.edu
Pennsylvania State University
United States of America

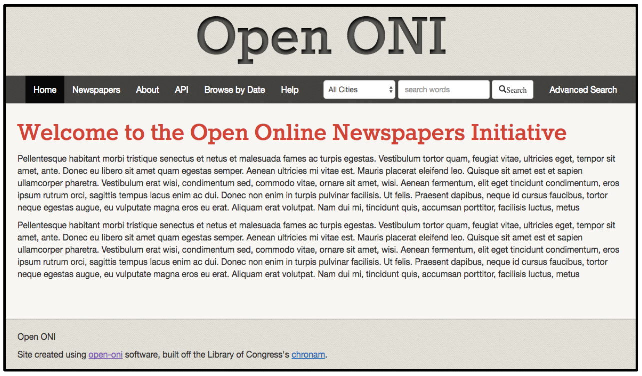
Sheila Rabun
srabun@clir.org
IIIF Consortium, United States of America

Greg Tunink
techgique@unl.edu
University of Nebraska, United States of America

The Open Online Newspaper Initiative (Open ONI) is an open source collaboration whose goal is to lower the entrance bar for libraries, archives, historical societies, and other cultural heritage institutions to display digital newspaper content. Open ONI was formed in response to a need for free, easily deployed, flexible, plug-and-play software that is useful for collections large and small, local and national.

Open ONI's code base was forked from the Library of Congress newspaper application, chronam (Library of Congress, 2016a). chronam was created to support the National Digital Newspaper Program (NDNP), a national project supported by the Library of Congress and the National Endowment for the Humanities, which seeks to digitize and add titles to a searchable, online collection at Chronicling America (Library of Congress, 2016d; 2016b). State entities often have newspaper content which they are unable to put online through Chronicling America and must seek a different solution to present their digitized periodicals (Library of Congress, 2016c; Center for Research Libraries Global Resources Network, 2015). The Library of Congress released the source code for chronam to help address this problem. The software currently available on GitHub requires skilled technical staff to customize the application and update related code packages, which can be a barrier for small institutions. Open ONI was born in 2015 when a group of librarians, project managers, and developers working on their own chronam installations gathered to discuss a friendly fork of the software. In the process of setting up, deploying, and customizing chronam for their own newspaper sites, Open ONI members had identified many shared interests, from fixing common bugs to building new features, and decided that pooling their resources and efforts to develop an application to fit the majority of their needs would benefit them all. Though we are currently working towards meeting our own implementation goals, we are keeping in mind how this might be applicable to others in the future.

Since beginning the initiative, Open ONI developers have made substantial changes and improvements to the initial chronam code. The web framework, Django, has been updated to the latest long term support version, and many supporting libraries and command line tools have been updated or replaced, when existing libraries and tools were deprecated or no longer available. Perhaps the largest improvement is the incorporation of the RAIS image server, an International Image Interoperability Framework (IIIF) compliant image server developed by the University of Oregon Library to deliver JP2s nearly as fast as chronam delivered TIFFs, but using a fraction of the RAM (Echols and Krech, 2016). RAIS is a 100% open source alternative to other JP2 image servers. With contributions from the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, and sponsored by the IIIF-Consortium, Open ONI offers IIIF compatible URLs to allow the substitution of other image servers. Beyond existing interoperability formats such as MARC and linked data, future plans could include IIIF manifests and other metadata formats/downloads, depending on developer time. Open ONI's developer environment relies on Docker containers to separate the different components of the software and provide a one-line setup command. The application is configurable through settings files and personalizable, with a default theme that users may copy to get started on their own look and feel. Additionally, Open ONI offers several plugins that users can incorporate into their site such as featured content, randomly selected newspaper pages, and a customizable map. A recent feature allows for defining copyright statements by publication and date range, and showing these statements directly below each page.



Welcome to the Open Online Newspapers Initiative

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis. Praesent dapibus, neque id cursus faucibus, tortor neque egestas augue, eu vulputate magna eros eu erat. Aliquam erat volutpat. Nam dui mi, tincidunt quis, accumsan porttitor, facilisis luctus, metus

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis. Praesent dapibus, neque id cursus faucibus, tortor neque egestas augue, eu vulputate magna eros eu erat. Aliquam erat volutpat. Nam dui mi, tincidunt quis, accumsan porttitor, facilisis luctus, metus

Open ONI

Site created using open-oni software, built off the Library of Congress's chronam.

Three states have plans to redeploy their newspaper sites with Open ONI in 2017: The University of Nebraska, University of Oregon, and Penn State University. Features that will be developed before all three launch their sites include extending the advanced search features for pages and newspaper titles, reinforcing the current test suite, updating the search engine from Solr 4 to Solr 6, and developing documentation to help organizations migrate from chronam to Open ONI.

Our poster will cover the above improvements of Open ONI, as well as the road map for future work. We will also have an instance of Open ONI available at the conference which will demonstrate the speed of the RAIS image server for JPG 2000s, default theme, new advanced search, and plugins. We hope that Open ONI offers organizations with newspaper collections of all types a reliable and customizable option for presenting and searching their collections. This poster presentation is also outreach to gain support and input from the archives, library, historical society, developer communities and others.

Bibliography

Center for Research Libraries Global Resources Network (2015). The ‘State of the Art’: A Comparative Analysis of Newspaper Digitization to Date. https://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf (accessed 24 October 2016).

Echols, J. and Krech, D. (2016). "RAIS Image Server." GitHub. https://github.com/uoregon-libraries/rais-image-server (accessed 24 October 2016).

Library of Congress (2016a). "chronam." GitHub. https://github.com/LibraryOfCongress/chronam (accessed 11 October 2016)

Library of Congress (2016b). "Chronicling America." http://chroniclingamerica.loc.gov/ (accessed 24 October 2016).

Library of Congress (2016c). “Content Selection.” http://www.loc.gov/ndnp/guidelines/selection.html (accessed 24 October 2016)

Library of Congress (2016d). "National Digital Newspaper Program." http://www.loc.gov/ndnp/ (accessed 24 October 2016).

Open Online Newspaper Initiative (2016). "Open ONI." GitHub. https://github.com/open-oni/open-oni (accessed 11 October 2016).

Go Queer

Maureen Engel
mengel@ualberta.ca
University of Alberta, Canada

Bamdad Aghilidehkordi
aghilide@ualberta.ca
University of Alberta, Canada

Kaitlyn Clafin
ekaitlyn@ualberta.ca
University of Alberta, Canada

Michaela Stang
mmstang@ualberta.ca
University of Alberta, Canada

Nailisa Tanner
nailisa@ualberta.ca
University of Alberta, Canada

This poster presents the work completed to date on *Go Queer*, a ludic, locative media experiment and experience that occurs on location, in the city, on the playful border between game and story, the present and the past, the queer and the straight, the normative and the *slant*. The app takes the city of Edmonton’s queer history as its text, and produces a locative, spatialized narrative of that history by displaying text, images, video and audio in place at the actual locations where they occurred, thus creating what Richardson and Hjorth (2014, 256) and call “the hybrid experience of place and presence.” The app invites its users to drift queerly through the city, discovering the hidden histories that always surround us, yet somehow remain just beyond our apprehension. The app compiles these traces into a media layer that augments quotidian city space, juxtaposing the past onto the present, creating a deep, queer narrative of place. By bringing together the physical navigation of the contemporary city with the imaginative navigation of its queer past, the app enacts a praxis that we characterize as a queer ludic traversal, one that renders the navigation itself as queer as the content that it presents. In so doing, the app produces the experience of *place*, in Lucy Lippard's (1997) formulation that

> Place is latitudinal and longitudinal within the map of a person's life. It is temporal and spatial, personal and political. A layered location replete with human histories and memories, place has width as well as depth. It is about connections, what surrounds it, what formed it, what happened there, what will happen there. (7)

This poster highlights four foundational questions of *Go Queer* that serve as base from which we research, write, and build:

Why a "game" for queer history?

Games have the unique affordance of putting players/readers into a first-person relationship with content, where they expect to actively interact with, and manipulate it. *Go Queer* seeks to generate the affective experience of queerness, to show and not just tell the queer history of the city, to collapse the distance between the present and the past. The game proceeds according to a procedural rhetoric that argues, in Sarah Ahmed's words (2006, 563), that "queers are differently 'oriented' to a spatial world that is the product of extended and extensive heteronormativity."

How can we encode a queer sense of place and co–presence in the contemporary city?

Thinking through the implications of "orientation" makes locative media a logical choice of format. It allows us to create a mimetic experience, where the player experiences the juxtaposition of queer space/time simultaneously with contemporary, normative space/time. The result of this juxtaposition is, we argue, a particularly powerful form of affect, where the player/user experiences the paradox of queer space: that it is always simultaneously both queer and not queer, slant and straight, persistent and fleeting.

How can we integrate crowd–sourced materials effectively and ethically?

Queer places are, by definition, sites of accretion, where stories, memories, and experiences are gathered. Queer place, in particular, is reliant on ephemeral histories, personal moments and memories. That *Go Queer* must integrate these personal archives is apparent, yet we must also be diligent about our ethical relationship to the material. How can we ethically solicit, verify, and evaluate community contributions to our history?

What technological structures will facilitate our goals?

The project requires both flexibility and stability. In order to offer the greatest flexibility to the interface, but the most stability to the actual content, we have decided on a custom server wrapped around a core GIS server. The GIS server will host the stable database; the custom wrapper will allow the flexibility to: a) customize and change the behaviour of the app as needed without modifying the underlying data structure; and b) input the primary materials into the database via a simple form-based interface.

We expect to have a functional beta version in place by DH 2017. A final production version should be launched in summer 2018.

Bibliography

Ahmed, S. (2006). "Orientations: Toward a Queer Phenomenology." *GLQ: A Journal of Lesbian and Gay Studies.* 12.4: 543-574.

Lippard, L. (1997). *The Lure of the Local*. New York: The New Press.

Richardson, I.,and Hjorth, L. (2014). "Mobile Games: From Tetris to Foursquare." In *The Routledge Companion to Mobile Media*, edited by Larissa Hjorth and Gerard Goggin, 256-266. New York: Routledge

Toward a Typology of Digital Thematic Research Collections

Katrina Simone Fenlon
kfenlon2@illinois.edu
University of Illinois at Urbana-Champaign
United States of America

Introduction

This paper considers an evolving genre of digital scholarship in the humanities, the thematic research collection, which is distinguished among other kinds of scholarly production as a collection of primary sources, gathered by scholarly effort and made available online to support research on a theme (Palmer, 2004). There are hundreds of such collections on the Web, ranging from well-known digital archives to small collections of historical or literary evidence within a thematic niche.

Despite recognition of the genre (e.g., Price, 2009; Flanders, 2014; and Thomas, 2015), we do not know enough about this mode of production, how it contributes to humanities discourse, or how it relates to systems of peer review, discovery, and long-term maintenance. The evolution of public-facing humanities scholarship, long-term access to collections, and the completeness of the scholarly record depend in part on a more systematic understanding of this and other emergent genres.

Through typological analysis, this research aims to build a foundation for rigorous study of thematic research collections. The goals of typology are to understand the breadth and variety of a genre, and identify unanticipated variations. We take up the following questions: What types of collections can we usefully distinguish, and what can these types and their characteristics reveal about the challenges and opportunities confronting the growth of digital scholarship in the humanities?

Method

We conducted a pilot survey of the digital humanities landscape to identify a set of resources meeting our definition. Sources for the survey included digital humanities centers, library publishing programs, tools and platforms for digital publishing, and scholarly collectives/peer review organizations. While not comprehensive of the digital humanities landscape, the survey produced a set of 98 diverse collections.

Our typology followed the formal process described by Kluge (2000):

1. Develop relevant analyzing properties. Properties reflect our intuitions about interesting differences between collections, within this context of scholarly work and use.
2. Group members by distinct combinations of properties.
3. Analyze meaningful relationships and construct types.
4. Repeat earlier steps if needed to accommodate collections that do not fit.

We iterated our analysis, refining our sense of properties and resultant types, until we were satisfied that our types speak to important and revelatory differences among collections.

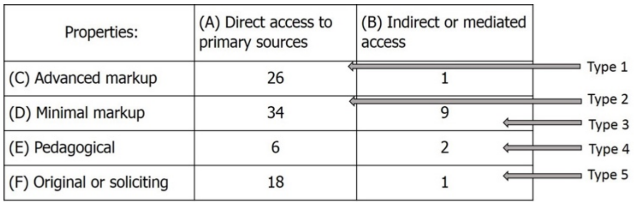
Analysis

Our proposed typology of thematic research collections relies on the following four properties of collections, which are basically determinative of their potential uses: (A) Whether primary sources are the main content of the collection, or are ancillary; (B) Whether the collection employs advanced markup, to enable use beyond basic keyword search; (C) Whether the collection's primary purpose is pedagogical; and (D) Whether the collection solicits, or actively engages in the collection of new or original evidence.

The 98 collections in our set resolve into 5 types, per different combinations of these properties. Figure 1 shows how types are derived from a matrix of properties, along with the number of collections that fall into each type. Figure 2 visualizes the types in a projected, three-dimensional property space.

We can briefly describe the types as follows:

- Type 1. Traditional collections with enabling markup: Marked-up (usually textual) primary sources constitute the main content of the collection, and are accessible directly by search and other functionalities.
- Type 2. Traditional collections without enabling markup: These are more heterogeneous in content, but primary sources are still directly accessible as such.
- Type 3. Data-centric or derivative-centric collections: While primary sources are a major component, they are not directly accessible as such. Rather, access is mediated by an analytic or interpretive layer, such as an interactive map or 3D model.
- Type 4. Pedagogical collections: They resemble one of the above types, but are distinguished by their intended purpose and audience.
- Type 5. Original or soliciting collections: They resemble one of the above types, but are distinguished by the scope and processes of their development (specifically, they are collecting new primary sources).



Properties:	(A) Direct access to primary sources	(B) Indirect or mediated access
(C) Advanced markup	26	1
(D) Minimal markup	34	9
(E) Pedagogical	6	2
(F) Original or soliciting	18	1

Type 1
Type 2
Type 3
Type 4
Type 5

Figure 1. Property matrix indicating numbers of collections in each type

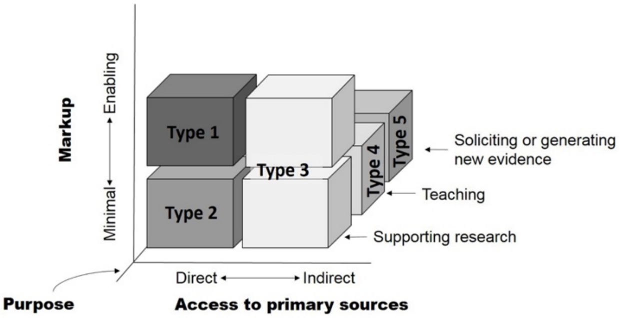


Figure 2. Types visualized in three-dimensional property space

Discussion

We hope the systematic identification of properties of a broad range of collections and preliminary types may serve as a foundation for ongoing study of how collections work, and how they may be served (or not) by systems of evaluation, discovery, and long-term maintenance. The full version of this poster details methods, properties, and types with vivid examples, and considers the implications of types for effective and ongoing access to collections.

Bibliography

Flanders, J. (2014). Rethinking Collections. In P. L. Arthur & K. Bode (Eds.), *Advancing Digital Humanities* (pp. 163–174). Palgrave Macmillan UK.

Flanders, J., & Jannidis, F. (2015). *Knowledge organization and data modeling in the humanities*. Providence, R. I.: Workshop on Knowledge organization and data modeling in the humanities, Brown University.

Flanders, J., & Muñoz, T. (2012). An Introduction to Humanities Data Curation. In *DH Curation: A Community Resource Guide to Data Curation for the Digital Humanities*. Champaign, IL.

Kluge, S. (2000). Empirically Grounded Construction of Types and Typologies in Qualitative Social Research. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(1).

Palmer, C. (2004). Thematic Research Collections. In *A Companion to Digital Humanities*. Blackwell Publishing.

Price, K. M. (2009). Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name? *Digital Humanities Quarterly*, *3*(3).

Thomas, W. G. (2015). The Promise of the Digital Humanities and the Contested Nature of Digital Scholarship. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A New Companion to Digital Humanities* (pp. 524–537). John Wiley & Sons, Ltd.

Modelling a Prosopography for the Roman Republic The Digital Prosopography of the Roman Republic Project

Luis Figueira
luis.figueira@kcl.ac.uk
King's Digital Lab
King's College London, United Kingdom

Miguel Vieira
jose.m.vieira@kcl.ac.uk
King's Digital Lab
King's College London, United Kingdom

In this abstract we describe the development of a web application for a prosopography of the Roman Republic, focusing on the data modelling, data harvesting and user interface. The web application offers fine-grained data about the elite of the Roman Republic, allowing the detailed study of its attested individuals, including familial composition, office-holding patterns, internal hierarchies, property and wealth.

The Digital Prosopography of the Roman Republic (DPRR) is an AHRC (Arts and Humanities Research Council) funded project, running from 2013 to May 2017. It is a collaborative project being developed by King's Digital Lab, together with the Classics and the Digital Humanities Departments from King's College London.

The project was developed to allow users easy access to complex data to enable them to answer questions such as:

- list the individuals who were praetors between 200-100 BCE with nomen Caecilius;
- list proconsuls in Gallia between 100-80 BCE;
- show all persons who were both consul and pontifex between 123 and 100 BCE (i.e. consuls who were also pontifices);
- visualise the composition of the the senate in the year 100 BCE;
- persons with a birth date and a death date between 250-100 BCE;
- women that died of violent death.

To preserve the richness of the data, while allowing answering research questions as above, the DPRR data model was designed to store the prosopographical data using the principles of the Factoid Model (Bradley and Short, 2002; Pasin and Bradley, 2012). This builds upon previous experience acquired developing similar projects (Prosopography of the Byzantine Empire, Clergy of Anglo-Saxon England Database, Prosopography of Anglo-Saxon England, Prosopography of the Byzantine World, People of Medieval Scotland, Making of Charlemagne's Europe).

The data model is organised around four main entities: *Person*, *PostAssertion*, *StatusAssertion* and *RelationshipAssertion*. The main entity, *Person*, stores information about each individual, such as names, gender, life dates, patrician indicator, etc. The *PostAssertion* entity stores data related to offices/posts held by individuals (such as praetor, consul, legate, etc), while the *StatusAssertion* stores senatorial and equestrian class information. The *RelationshipAssertion* represents personal relationships (brother of, son of, etc), connecting two different individuals within the database. All these entities are linked, as represented in Figure 1. All Assertion models are linked to a *SecondarySource* entity with details about the bibliographical references supporting the data.

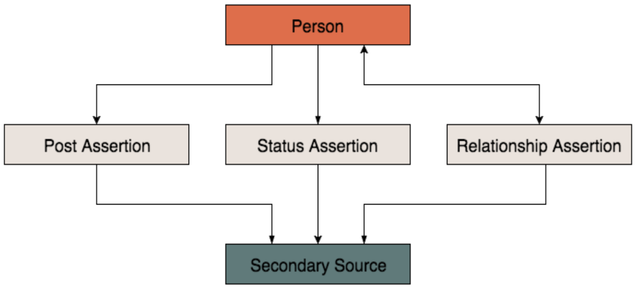


Figure 1 – Simplified database model

The data for the project has been harvested from multiple sources – mainly from Broughton's Magistrates of the Roman Republic (1951-86), Ruepke's Fasti Sacerdotum (2005) and Zmeskal's Adfinitas (2009). The data was automatically loaded using scripts developed for each source (Fig. 2). There's also a data editing interface, which was mostly used to make corrections to the automatically loaded data, as well as adding new information when needed.

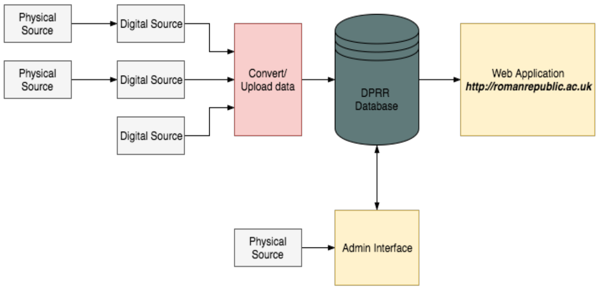


Figure 2 – Data workflow

Besides loading and entering data directly into the database, the project also has rules to infer new information about personal relationships and senators. An example is, the application creates a Senator object for a person that held a specific post in a given year, such as consul, praetor, aedile, tribunus plebis, censor, princeps senatus.

The database, which stores all this information in an easily searchable format, is fronted by a web application (see the public version of the website). This website allows the public to explore the data using a faceted search interface with different filters that can be applied. This guides the user to achieve answers to the research questions above. By selecting individual results, it is possible to see all the information about a person, including visualisations of their relationship networks (Fig. 3).

Visualisations of the senate composition on a yearly basis are also being created. We are currently working towards integrating the database with Linked Data technologies as part of the Standard for Networking Ancient Prosopographies (see the SNAP:DRGN project).

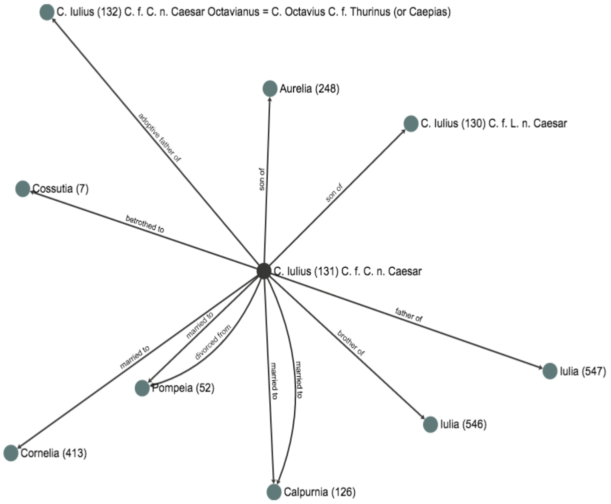


Figure 3 – Network visualisation for *C. Iulius (131) C. f. C. n. Caesar*

The project is built using open source tools and technologies, mainly the Django application framework with a PostgreSQL database. The search interface is implemented using the Solr search engine together with the django-haystack search package for Django. The visualisations are created with *d3plus* and *linkurious*. The project source code is also available as open source in a GitHub repository.

Our poster will illustrate the model that supports this digital resource, the methods used to retrieve the data and populate it, as well its usage for data exploration and visualisation.

Bibliography

Bradley, J., Short, H. (2002) Using formal structures to create complex relationships: the prosopography of the Byzantine Empire - a case study. *Resourcing Sources Prosopographica et Geneologica* 7

Pasin, M, and Bradley, J. (2013) Factoid-based prosopography and computer ontologies: towards an integrated approach *Digital Scholarship in the Humanities* 30.1 (2013): 86-97.

Re(a)d Wedding: A Comparative Discourse Analysis of Fan Responses to Game of Thrones

Eric Forcier
School of Information Studies
Charles Sturt University, Australia

It is no exaggeration to say that HBO's *Game of Thrones* is more than just a television series or a successful brand: it is a transmedia system in the sense first used by Marsha Kinder (1991) and popularized by Henry Jenkins (2006), in which media-hopping networks of intertextualities extend the "storyworld" of an original production. Now spanning six seasons and 60 episodes, with an average global viewership (from its most recent season) of 25.1 million viewers per episode (Shepherd, 2016), it has spawned five video games, a graphic novel adaptation, several companion books, two rap albums, a 28-city orchestral tour, a wide variety of tabletop games, toys, merchandise and mobile apps, and countless podcasts, fanfics and other fan-based creations. Given the volume of content this represents, it is easy to forget that the television series itself is an adaptation of a book series with a pre-existing fandom. As such, the *Game of Thrones* storyworld represents a remarkably rich and challenging environment for fans old and new, who must negotiate an increasingly complex network of paratexts and intertexts in order to fully engage with its narratives.

In this sense, fans of the series represent an emerging model for cultural consumption that should be carefully explored. Transmedia systems, like that exemplified by *Game of Thrones*, are becoming increasingly prevalent (e.g., *Star*

Wars, Harry Potter, The Walking Dead, the Marvel Cinematic Universe, etc); these systems demonstrate, in microcosm, the global challenge of managing the fire-hose flow of information in contemporary postdigital society. The study of how people, as fans, access and manage information within a transmedia system provides valuable insight that contributes not only to practitioners and scholars of the media industry, but to the wider context of cultural studies, by offering findings on this new model of the fan as consumer and information-user. For us, as digital humanists, defining the "transmedia fan" is of particular relevance as we seek to understand contemporary social and cultural transformations engendered by digital technologies.

Methodology

As a first step in defining the "transmedia fan", the current project undertakes a comparative discourse analysis of online conversations of *Game of Thrones* fans. One of the most dramatic plot developments in the source material (Martin, 2000) was adapted to the screen in the penultimate episode of the third season, "The Rains of Castamere" (Benioff & Weiss, 2013). Readers of the book series had long anticipated and dreaded the events of the "Red Wedding", while fans of the show unfamiliar with Martin's narrative were largely taken unawares by the pivotal episode.

Since the television series' inception, writers at *The AV Club* have written two critical reviews for each episode: one for viewers familiar with the books (i.e., "Experts") and one for viewers unfamiliar with the books and averse to "spoilers" (i.e., "Newbies"). What results are two completely separate reviews of "The Rains of Castamere" which in turn document the fans' reactions to the episode in the form of user comment threads: one comment thread where fans were expected to be shocked by the outcome of the episode and one comment thread where fans had hotly anticipated it.

As a pilot project, the current work takes the content of both comment threads—a corpus of approximately 5,600 comments—and analyzes each thread separately using a qualitative coding method aligned with constructivist grounded theory (Charmaz, 2006). Through this analysis, a categorization of themes emerges illustrating tactics for negotiating intertexts and paratexts unique to each group of fans. These themes fall under two broad categories: sentimental negotiation (i.e., emotional responses) and tactical negotiation (i.e., cognitive, or reasoned responses). A comparison of categories and sub-categories between both groups provides preliminary findings to support an emergent model, or models, of the "transmedia fan".

Conclusion

The present research represents a first step in exploring the impact of transmedia systems, as exemplified by *Game of Thrones*, through the study of fans. The question posed by this research is, fundamentally, an examination of how the problem of "access" is framed in postdigital society from the perspective of the consumer. Future research should explore the negotiation tactics observed in transmedia fans using the principles of De Certeau's (1984) everyday life practice, in order to extend its application to the broader context of modern-day consumers. The current study will contribute to the development of further qualitative and quantitative research that will more clearly define the information behaviors of the transmedia fan. This project is of relevance to researchers in media studies, fan studies, information studies and digital humanities

Bibliography

Benioff, D. and Weiss, D.B (2013). "The Rains of Castamere", Dir. David Nutter, aired June 2, 2013. In *Game of Thrones* [Television Series], Prod. David Benioff and D. B. Weiss. HBO.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis.* Thousand Oaks, CA: Sage.

de Certeau, M. (1984). *The Practice of Everyday Life*. Trans. Steven Randall. Berkeley, CA: University of California Press.

Jenkins, H. (2006). *Convergence Culture: Where Old and New Media Collide.* New York: New York University Press.

Kinder, M. (1991). *Playing with Power in Movies, Television, and Video Games: From Muppet Babies to Teenage Mutant Ninja Turtles*. Berkeley, CA: University of California Press.

Martin, G. R. R. (2000). *A Storm of Swords*. New York: Voyager Books.

Shepherd, J. (2016). Game of Thrones season 6 ratings: Show brought in 25.1 million viewers on average per episode. July 19, 2016. *The Independent.* Retrieved from http://www.independent.co.uk/arts-entertainment/tv/news/game-of-thrones-season-6-ratings-show-brought-in-251-million-viewers-on-average-per-episode-a7144261.html

OBSERVATÓR!O2016

Julia Gianella
julia@impa.br
Instituto de Matemática Pura e Aplicada, Brazil

Luiz Velho
lvelho@impa.br
Instituto de Matemática Pura e Aplicada, Brazil

This poster focuses on OBSERVATÓR!O2016, a web-based platform for collecting, structuring and visualizing the online response to Rio-2016 from content shared on Twitter. This project was developed at the Vision and Computer Graphics Laboratory and is based on two cross related research lines. First, the conception and design of the OBSERVATÓR!O2016 website, which provides a space to explore comments and images about the Olympics through structured visualizations. Second, the ongoing deployment of research regarding the application of a digital method to explore large sets of images. The goal is to highlight how we use *Deep Learning* applications - such as automatic image classification - and visualization techniques to enhance discoverability and expression of subject features within an extensive image collection.

Method: *Deep Learning* and mosaic visualization

OBSERVATÓR!O2016 collected around 1 million tweets from April 18th to August 25th, 2016. In other to gather different perspectives on the debate about the Olympics we created seven Twitter search queries. The data was presented in eight interactive visualizations:

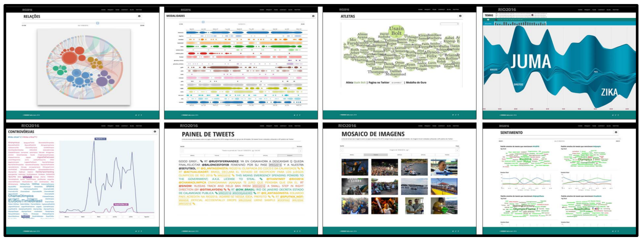


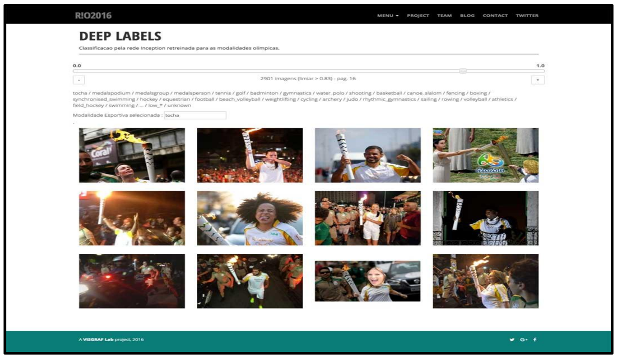
Figure 1. Visualizations

Approximately 180,000 of the collected tweets included unique images. The production of large sets of digital images inaugurates new avenues for researchers interested in the human creative practice. In this sense, the investigation of Lev Manovich on digital methods to study visual culture is quite relevant. With this in mind, we explored Rio-2016 images through *Deep Learning* approaches. As the investigation is currently ongoing, we will report the research process of a single task, which resulted in the *Torch Mosaic* visualization.

During the pre-Olympics, it became evident that many of the images gathered by our query scripts were related to the Olympic torch relay and depicted the iconic object. Part of these images were accompanied by texts that mentioned the torch, but not all. In addition, some tweets mentioning the torch relay incorporated images that didn't depict the object. In other words, text analysis alone was not sufficient to detect a set of images containing the torch.

Thus, we referred to a *Deep Learning* approach to recognize the Olympic torch in our database. The field of visual pattern recognition has been recently improved by the efficient performance of *Convolutional Neural Networks* (CNN). In 2012, the work of Krizhevsky et al. on training a deep convolutional neural network to classify the 1.2 million images in the *ImageNet* LSVRC-2010 contest into 1000 different classes had a substantial impact on the computer vision community.

More recently, and thanks to Google, computer vision tasks such as image classification have become more accessible and applicable. That's because the company released last year their open source software library *TensorFlow.* This library runs code for image classification on *Inception-v3* CNN model, which can be retrained on a distinct image classification task (this quality is referred to as *transfer learning*). By creating a set of training images, it is possible to update the parameters of the model and use it to recognize a new image category. That said, we retrained the network by showing it a sample of 100 manually labeled images containing the torch. Finally, the retrained network ran over our database and returned a set of images with their corresponding confidence score for the new category.



DEEP LABELS

Classificacao pela rede Inception retreinada para as modalidades olimpicas.

0.0 1.0

2901 imagens (limiar > 0.83) - pag. 16

tocha / medalspodium / medalsgroup / medalsperson / tennis / golf / badminton / gymnastics / water_polo / shooting / basketball / canoe_slalom / fencing / boxing / synchronised_swimming / hockey / equestrian / football / beach_volleyball / weightlifting / cycling / archery / judo / rhythmic_gymnastics / sailing / rowing / volleyball / athletics / field_hockey / swimming / ... / low_* / unknown

Modalidade Esportiva selecionada : tocha

A VISGRAF Lab project, 2016

Figure 2. Confidence score over 83%

Until June 25th, around 1500 images with over 85% confidence score for the Olympic torch category had been classified by our network. We used them to create a mosaic visualization that can be zoomed and panned. The mosaic idea is that, given an image (target image), another image (mosaic) is automatically build up from several smaller images (tile images). To implement the mosaic, we used a web-based viewer for high-resolution zoomable images called *OpenSeadragon*.

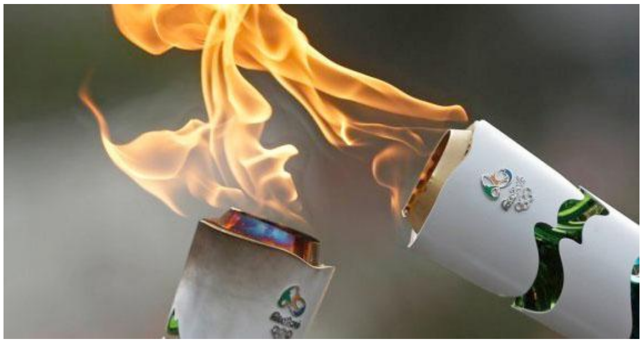


Figure 3. The target image

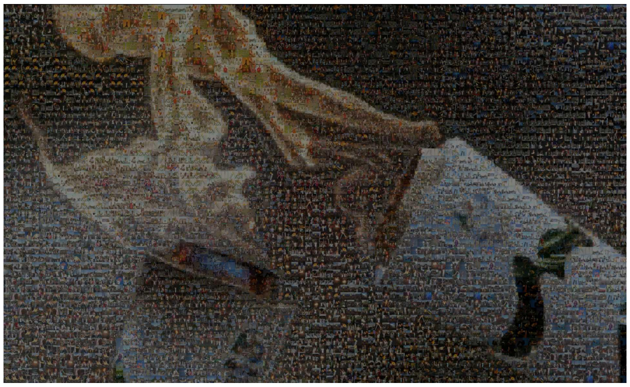


Figure 4. The mosaic

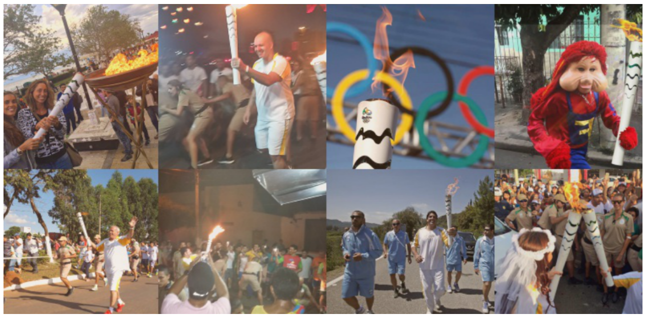


Figure 5. Tile images

The organization nature of the mosaic visualization is mainly aesthetic. Nevertheless, zooming and panning the mosaic allows the user to explore a wide variety of views, and to discover image details and surprises such as the spoof picture of *Fofão*, a Brazilian fictional character, carrying the torch.

For DH2017 poster session, we expect to present the results for another subject feature - the sporting disciplines - and visualisation technique - the *videosphere* - we are working on at the moment. In addition, we plan to discuss with participants some possible scenarios in which *Deep Learning* models could be applied to help image collections become more discoverable and expressive.

Bibliography

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems,* pp. 1097-1105.

Manovich, L (2012). "How to Compare One Million Images?". In Berry, D. (ed), *Understanding Digital Humanities.* Palgrave, pp. 249-278.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). "Going deeper with convolutions". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

What's in a word? Exploring words and their usage in the *Dictionnaire Vivant de la Langue Française*

Clovis Gladstone
clovisgladstone@gmail.com
ARTFL Project, University of Chicago
United States of America

Charles Cooney
chu.cooney@gmail.com
ARTFL Project, University of Chicago
United States of America

Tim Allen
timothy.d.allen@gmail.com
ARTFL Project, University of Chicago
United States of America

Originally funded by a startup grant from the National Endowment for the Humanities, *Le Dictionnaire Vivant de la Langue Française* (DVLF) is an experimental approach to dictionary compilation that aims to push the boundaries of what typical dictionaries offer. It is being developed as an interactive, community-oriented alternative to traditional methods of French lexicography, bringing together user-submitted definitions and definitions from standard French dictionaries (multiple editions of the *Dictionnaire de l'Académie Française, Dictionnaire de Littré...*), synonyms, pronunciation, as well as a wealth of information on word usage through the use of various computational methods. Our poster will show the DVLF's functionality, as well as demonstrate its originality.

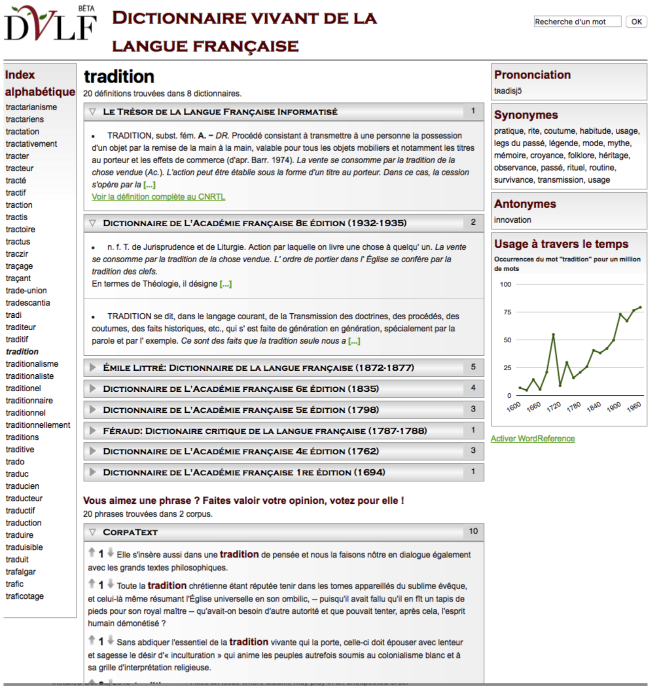
Fundamentally, the most important aspect of the DVLF is that it aims to create an environment in which its user community rates, critiques, and adds to the collection's resources as it sees fit. As such, we have made a conscious effort to facilitate community engagement by allowing users to contribute to most sections of our website: definitions, usage examples, synonyms and antonyms. To date, users have added or submitted definitions/examples for hundreds of words, including **alexithymie**, **digiscopie**, **foucade**, and **nyctalope**. The new version of our site (released in January 2017) has been further enhanced to enable such engagement thanks to more visible links inviting users to contribute content, as well as a new responsive interface designed to work equally well on mobile devices and desktop computers.

In order to attract and engage the largest possible user community, the DVLF is entirely free and open source, and requires no registration. Internet users at large can access the site's resources and contribute material. The DVLF thus tries to adapt to changing word usage and incorporate neologisms so that users will be able to select the word senses and usage examples that they feel are most consistent with contemporary usage. In this manner, the DVLF mirrors the social and evolving nature of language, expanding upon the dictionary's traditionally normative role by giving French speakers and learners access to lexicographic tools so that they might interact with the evolving meaning of words and determine their own understanding of the language.

Over the last year, ARTFL has been working to develop and improve the DVLF. We noticed from statistics gathered from Google Analytics that our community of users was very diverse, accessing our website from Morocco, Tunisia, Canada, and many other francophone and non-francophone countries. This led us to focus a large part of our effort on diversifying content, including usage examples from a wider range of francophone sources in order to provide more coherent descriptions of emerging word usage from Francophone communities around the world, and therefore attract a larger global user base.

Over the course of this redevelopment effort, we decided to rewrite the entire codebase given the various reliability and performance issues we had experienced, switching from a Python-only infrastructure to a Go/Javascript environment. We have also worked to add additional contextual information to our content, starting with the most frequent collocates of any given word based a corpus of over 7,000 texts of French literature from the Middle-Ages to the late 20th century. Additionally, following the lead of recent work in computational linguistics and natural language processing, we also now provide the nearest neighbors based on the computation of word vectors using a word-embedding technique called Swivel. While we could have used alternative algorithms such as word2vec or Glove, we decided to use Swivel because it does not solely rely only on word co-occurrence to construct vectors. The benefit of Swivel's approach is that it can yield similar words that do not actually occur together in the corpus (see Shazeer et al, 2016), which we thought was a valuable approach given that our dictionaries span across multiple centuries. To generate this word vector model, we used the same 7,000 texts and then extracted the top 20 words for every headword in our dictionary index, which is now displayed on the new version of our site.

Appendix: Screenshots



DVLF BÊTA DICTIONNAIRE VIVANT DE LA LANGUE FRANÇAISE

Recherche d'un mot OK

Index alphabétique

tractarianisme
tractariens
tractation
tractativement
tracter
tracteur
tracté
tractif
traction
tractis
tractoire
tractus
traczir
traçage
traçant
trade-union
tradescantia
tradi
traditeur
traditif
tradition
traditionalisme
traditionaliste
traditionel
traditionnaire
traditionnel
traditionnellement
traditions
traditive
trado
traduc
traducien
traducteur
traductif
traduction
traduire
traduisible
traduit
trafalgar
trafic
traficotage

tradition

20 définitions trouvées dans 8 dictionnaires.

▽ LE TRÉSOR DE LA LANGUE FRANÇAISE INFORMATISÉ 1

- TRADITION, subst. fém. **A.** – *DR.* Procédé consistant à transmettre à une personne la possession d'un objet par la remise de la main à la main, valable pour tous les objets mobiliers et notamment les titres au porteur et les effets de commerce (d'apr. Barr. 1974). *La vente se consomme par la tradition de la chose vendue* (Ac.). *L'action peut être établie sous la forme d'un titre au porteur. Dans ce cas, la cession s'opère par la* [...]

Voir la définition complète au CNRTL

▽ DICTIONNAIRE DE L'ACADÉMIE FRANÇAISE 8E ÉDITION (1932-1935) 2

- n. f. T. de Jurisprudence et de Liturgie. Action par laquelle on livre une chose à quelqu' un. *La vente se consomme par la tradition de la chose vendue. L' ordre de portier dans l' Église se confère par la tradition des clefs.*
En termes de Théologie, il désigne [...]

- TRADITION se dit, dans le langage courant, de la Transmission des doctrines, des procédés, des coutumes, des faits historiques, etc., qui s' est faite de génération en génération, spécialement par la parole et par l' exemple. *Ce sont des faits que la tradition seule nous a* [...]

▶ ÉMILE LITTRÉ: DICTIONNAIRE DE LA LANGUE FRANÇAISE (1872-1877) 5

▶ DICTIONNAIRE DE L'ACADÉMIE FRANÇAISE 6E ÉDITION (1835) 4

▶ DICTIONNAIRE DE L'ACADÉMIE FRANÇAISE 5E ÉDITION (1798) 3

▶ FÉRAUD: DICTIONAIRE CRITIQUE DE LA LANGUE FRANÇAISE (1787-1788) 1

▶ DICTIONNAIRE DE L'ACADÉMIE FRANÇAISE 4E ÉDITION (1762) 3

▶ DICTIONNAIRE DE L'ACADÉMIE FRANÇAISE 1RE ÉDITION (1694) 1

Vous aimez une phrase ? Faites valoir votre opinion, votez pour elle !

20 phrases trouvées dans 2 corpus.

▽ CORPATEXT 10

↑ 1 ↓ Elle s'insère aussi dans une **tradition** de pensée et nous la faisons nôtre en dialogue également avec les grands textes philosophiques.

↑ 1 ↓ Toute la **tradition** chrétienne étant réputée tenir dans les tomes appareillés du sublime évêque, et celui-là même résumant l'Église universelle en son ombilic, -- puisqu'il avait fallu qu'il en fît un tapis de pieds pour son royal maître -- qu'avait-on besoin d'autre autorité et que pouvait tenter, après cela, l'esprit humain démonétisé ?

↑ 1 ↓ Sans abdiquer l'essentiel de la **tradition** vivante qui la porte, celle-ci doit épouser avec lenteur et sagesse le désir d'« inculturation » qui anime les peuples autrefois soumis au colonialisme blanc et à sa grille d'interprétation religieuse.

Prononciation

tʀadisjɔ̃

Synonymes

pratique, rite, coutume, habitude, usage, legs du passé, légende, mode, mythe, mémoire, croyance, folklore, héritage, observance, passé, rituel, routine, survivance, transmission, usage

Antonymes

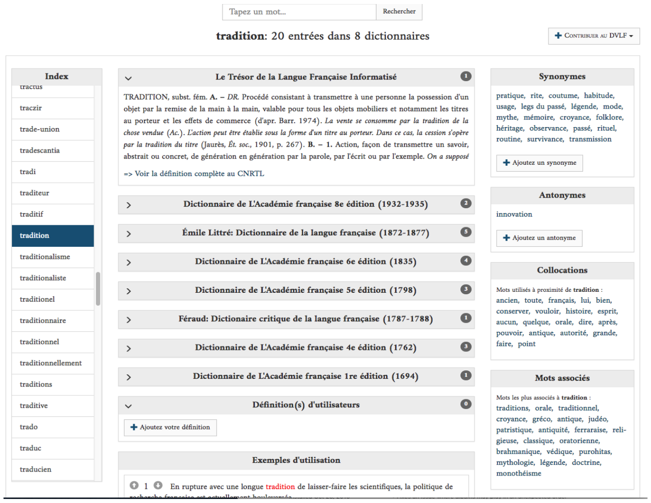
innovation

Usage à travers le temps

Occurrences du mot "tradition" pour un million de mots

Activer WordReference

Figure 1. Old version of the site (retired on January 2017)



Tapez un mot... Rechercher

tradition: 20 entrées dans 8 dictionnaires

✚ CONTRIBUER AU DVLF ▾

Index

tractus
traczir
trade-union
tradescantia
tradi
traditeur
traditif
tradition
traditionalisme
traditionaliste
traditionel
traditionnaire
traditionnel
traditionnellement
traditions
traditive
trado
traduc
traducien

Le Trésor de la Langue Française Informatisé 1

TRADITION, subst. fém. **A.** – *DR.* Procédé consistant à transmettre à une personne la possession d'un objet par la remise de la main à la main, valable pour tous les objets mobiliers et notamment les titres au porteur et les effets de commerce (d'apr. Barr. 1974). *La vente se consomme par la tradition de la chose vendue* (Ac.). *L'action peut être établie sous la forme d'un titre au porteur. Dans ce cas, la cession s'opère par la tradition du titre* (Jaurès, *Ét. soc.*, 1901, p. 267). **B.** – **1.** Action, façon de transmettre un savoir, abstrait ou concret, de génération en génération par la parole, par l'écrit ou par l'exemple. *On a supposé*

=> Voir la définition complète au CNRTL

› **Dictionnaire de L'Académie française 8e édition (1932-1935)** 2

› **Émile Littré: Dictionnaire de la langue française (1872-1877)** 5

› **Dictionnaire de L'Académie française 6e édition (1835)** 4

› **Dictionnaire de L'Académie française 5e édition (1798)** 3

› **Féraud: Dictionaire critique de la langue française (1787-1788)** 1

› **Dictionnaire de L'Académie française 4e édition (1762)** 3

› **Dictionnaire de L'Académie française 1re édition (1694)** 1

⌄ **Définition(s) d'utilisateurs** 0

✚ Ajoutez votre définition

Exemples d'utilisation

1 En rupture avec une longue tradition de laisser-faire les scientifiques, la politique de

Synonymes

pratique, rite, coutume, habitude, usage, legs du passé, légende, mode, mythe, mémoire, croyance, folklore, héritage, observance, passé, rituel, routine, survivance, transmission

✚ Ajoutez un synonyme

Antonymes

innovation

✚ Ajoutez un antonyme

Collocations

Mots utilisés à proximité de **tradition** :
ancien, toute, français, lui, bien, conserver, vouloir, histoire, esprit, aucun, quelque, orale, dire, après, pouvoir, antique, autorité, grande, faire, point

Mots associés

Mots les plus associés à **tradition** :
traditions, orale, traditionnel, croyance, gréco, antique, judéo, patristique, antiquité, ferraraise, religieuse, classique, oratorienne, brahmanique, védique, purohitas, mythologie, légende, doctrine, monothéisme

Figure 2. New version of the site (went live on January 2017)

An Open, Reproducible Method for Teaching Text Analysis with R

Tassie Gniady
ctgniady@iu.edu
Indiana University, United States of America

Eric Wernert
ewernert@iu.edu
Indiana University, United States of America

Over the past year and a half, the Cyberinfrastructure for Digital Humanities (CyberDH) Group at Indiana University has been developing an open instructional workflow for text analysis that aims to build algorithmic understanding and basic coding skills before scaling up analyses (Gniady et al., 2017). We have chosen to bootstrap in R, a high level and high productivity language, with methods that are open, repeatable, and sustainable. The aim is to provide code templates that can be adapted, remixed, and scaled to fit a wide range of text analysis tasks. This poster presents our approach to teaching computational text analysis and a representative hypothetical case study in which two different users are able to start with the same corpus and adapt code to achieve very different end results in a way not currently possible with black box tools.

This paradigm is fundamentally different from that currently practiced by many in the digital humanities. Blackboxed tools with GUIs that hide computation are very popular for introducing new practitioners of text analysis in the digital humanities to basic algorithms and outputs. In 2012, AntConc was downloaded 120,000 times by users in 80 different countries (Anthony, 2014). Voyant 1.0 had 113 sites linking to it actively in 2012 (Sinclair and Rockwell, 2013) and the week Voyant 2.0 was released the server went down multiple times from excess traffic (@VoyantTools, 2016). However, one of its default corpora is the Shakespearean dramas, with speaker names and stage directions. ((Sinclair and Rockwell, 2016). The inclusion of speaker names skews all algorithms related to frequency counts of characters (e.g. word clouds), which a new user may not even think to take into account. Using AntConc's concordance tool with a Shakespearean corpora including speaker names gives an idea of when a character speaks **and** when a character is mentioned, but this conflation might not jump out at a new user. If anything, we suggest learning about algorithms **first** and then moving up to black-box tools when one has the means to critique them.

Having looked at popular "plug-and-play" tools for corpora visualization, it becomes evident that even simple visualizations can lead to inaccurate results if the user is not thinking through how a corpus is being processed to produce a result. We believe that if the user understands how the algorithm is generating visualizations, they can contribute more meaningfully to critiques of sophisticated algorithms when partnered with programmers or even go on to bootstrap themselves with awareness of their domain's particular caveats. Thus, we advocate teaching humanists the basics of coding to create **conversant programmers** similar to the methodology behind Matthew Jockers' *Text Analysis with R for Students of Literature*, but with a slightly slower ramp up. To this end we have a three-step process of introducing R: web-deployed Shiny apps, highly marked up RNotebooks, and lightly commented RScripts, both in "regular" and higher performance versions. All are available for download on Github (with associated sample data from Shakespeare and Twitter) (CyberDH Team, 2017). We hope that this simpler bootstrapping method that mixes code and explanation, pedagogy and self-driven inquiry, will be of use to those looking to onramp new practitioners who may go on to partner with programmers if needed or to remix available code to look at their own knowledge domain.

Bibliography

Anthony, L. (2016). Antconc 3.4.4. Software. http://www.laurenceanthony.net/software/antconc/.

Gniady, T. Thomas, G. and Kloster, D. (2017). *Text Analysis Github Repository*. https://github.com/cyberdh/Text-Analysis.

Jockers, M. (2014). *Text Analysis with R for Students of Literature*. New York: Springer International Publishing.

Sinclair, S. and Rockwell, G. (2013). "Voyant Notebooks: Literate Programming, Programming Literacy." *Digital Humanities 2014: Conference Abstracts.* Nebraska-Lincoln: http://dh2013.unl.edu/abstracts/ab-295.html.

Sinclair, S. and Rockwell, G. (2016). *Voyant Tools*. http://voyant-tools.org/.

@VoyantTools. Twitter. 8 April 2016.

Information Infrastructure of Linked Data for Promoting Integrated Studies of Cultural and Research Resources

Makoto Goto
m-goto@rekihaku.ac.jp
National Museum of Japanese History, Japan

Ayako Shibutani
shibutani@rekihaku.ac.jp
National Museum of Japanese History, Japan

The National Museum of Japanese History (NMJH) is using digital technology to establishing its core research project, Constructing Integrated Studies of Cultural and Research Resources. This project offers unique insight into a variety of studies and Japanese historical resources through multidisciplinary collaboration with universities, museums, and other institutes. This paper introduces the initial application of Linked Data and IIIF (International Image Interoperability Framework).

Researchers working in the humanities and sciences with a focus on history need to collaborate with their peers in relevant fields to produce more diversity and substance in academic data. The plan for a new academic field is to construct a "research circulation model" that links academic studies and resource information through advanced cooperative studies. Although institute resources encompass a wide variety of materials, the poor quality of the current infrastructure has hampered researchers' access to original data and convenience in producing the primary evidence for research results.

The number of researchers specializing in the study of China, Korea, and Taiwan is increasing due to the rapid growth in the networks of historical resource information in these countries. In contrast, the number of their peers in Japan is decreasing, in part because of Japan's non-digitized and partially closed systems. This trend is particularly apparent in the younger generation, which is accustomed to using the internet. In our poster, we show an example of our collaborative studies with relevant local and international institutes. The name authorities and concept label authorities in the translation are related to the entities mentioned in the RDF (Resource Description Framework). In addition to preservation of data in the case of disaster, the poster presents our development of a support system in the affected areas and an advanced sharing infrastructure.

RDF is a de facto standard since it is easy to add and edit metadata. It also enables the addition of multiple metadata to a single data element. Giving each resource its own address produces a permanent link to permanent data. This is an advantage for any researcher or institute that uses the data as an academic resource. Linking one resource to multiple relevant resources enables one to find information resources consecutively, as well as resources followed by historical studies. Research results must refer to original resources. In addition to fundamental data, division of systems and RDF can share complex data and advanced applications. Our RDF-based prototype is currently being tested, but it can connect internal and external databases by linking their URLs.

To achieve the current results from resources in Japan, we also constructed an initial application of Linked Data and IIIF. It has five features. First, each resource has its one own address, which enables access to its resources from Google and other search engine results. Second, the application can access resources that follow another resource, such as a resource ☾an institute ☾another resource. Third, for smart device applications, spatial information is added to the resources. Fourth, the system and the data are separate. If a system is replaced or updated in the future, the data should remain usable. If an institute provides infrastructure data, other institutes or researchers can use the data to create applications. In doing so, the access path to the institute needs to be confirmed. Finally, RDF data are applied to the database.

We are currently implementing all the functions and testing for real data, including 150,000 triples of our museum collection, 20,000 triples of related de facto data, and approximately 500 IIIF images. We plan to integrate the data sets of other Japanese universities by August 2017, and the application will be fully launched at the beginning of 2018.

Here, we discuss links using Ukiyo-e of Japan. The Ukiyo-e material enables access to information concerning its holding institute as well as access to other collections. Its spatial information is also added to the resources. If, for example, Mt. Fuji is drawn on the Ukiyo-e, this application can show its spatial information. The application also connects to image data using IIIF (providing both Image API and Presentation API) from RDF data.

We are now constructing a system to connect information about the Ukiyo-e artist to information about his work. Some of these links include metadata, such as FOAF (Friend of a Friend) in the information about people, and other links use vocabulary defined by our project.

Complex Network Visualisation for the History of Interdisciplinarity: Mapping Research Funding in Switzerland

Martin Grandjean
martin.grandjean@unil.ch
University of Lausanne, Switzerland

Pierre Benz
pierre.benz@unil.ch
University of Lausanne, Switzerland

Thierry Rossier
thierry.rossier@unil.ch
University of Lausanne, Switzerland

Introduction

In Switzerland, the panorama of scientific research is deemed to be deeply affected by language barriers and strong local academic identities. Is this impression confirmed by data on research projects? What are the factors that best explain the structure of scientific collaborations over the last forty years? Do linguistic regions (Switzerland is divided into three principals) or local academic logics really have an impact onto the mapping of research collaborations and to what extend are they embedded in disciplinary, historical and generational logics?

We focus on the very large database of the Swiss National Science Foundation (SNSF), the principal research funding agency in Switzerland, which lists all the 62,000 projects funded between 1975 and 2015. While scientometric studies generally focus on measuring work – and financial – performance, we aim to raise awareness on pursuing a socio-history analyse of Swiss academic circles by crossing the SNSF data with a prosopographic database of all Swiss university professors in the twentieth century provided by the Swiss Elite Observatory (OBELIS). Beyond the interest for the history of science and universities, we explore the noteworthy technical challenge of a network analysis of nearly 88,000 researchers and more than a million of collaborations.

By combining those two databases, we measure the temporality and spatiality of academic collaborations, i.e. to define a way to deal with the volume of information in order to provide not only a global vision but also to enable a fine processing of personal trajectories.

Sources

The SNSF database has been placed under an Open Data licence in spring 2016. Called "P3" for "Projects, People, Publications", it contains detailed information on all the projects funded since 1975 (around 500 per year in the beginning, almost 3,000 per year today, see Fig.1), as well as the whole list of people involved in the projects. The database can sometimes be incomplete about the discipline and institutional affiliation of individuals, since it depends directly on the project submission interface where some fields may be left empty. Thus, this gap is partly offset by the junction with the Swiss professors database that provides systematic data on Swiss professors. Thus, the projects are classified according to a standard tree of scientific disciplines.

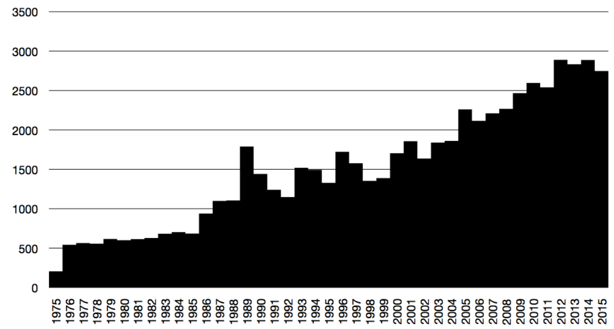


Figure 1. Evolution of the number of projects funded by SNF annually from 1975 to 2015.

Methodology

We are interested in the 2006-2015 period, ten years during which 25,000 projects involving 45,000 people produce a graph of more than 350,000 edges. On the one hand, this short periodization allows us to confront our assumptions to our data before analysing the full corpus. On the other hand, it helps to test the effectiveness of our tools and the interoperability of the two databases to prepare a complete and longitudinal modelling.

We therefore extracted a 2-mode network of people and projects from the database and then projected it into a 1-mode network of people only (the nature of the link being to be affiliated as collaborators to the same research project). If usually a relatively simple task, the transformation of a 2-mode graph into a 1-mode graph is here greatly complicated by the mass of information to process: when the graph matrix contains billions of positions, most softwares are reaching their limits. We will then divide the dataset into smaller units (here, transforming the network year after year helps make it bearable to a standard processor).

Analysis and Visualisation

The topography of the network obtained for 2006-2015 (Fig.2) is quite remarkable. The center of the network is not, as it is often, the densest region, which would have meant that a single discipline or field of study was likely to play a role of interface between others. Instead, we observe an almost circular distribution of individuals, recalling other "science maps" based on the organization of institutions of bibliometric analysis (Rafols et al. 2010). Data visualization, and in particular the representation of complex networks, is not an end in itself but a tool for questioning the structure of the dataset (Grandjean 2015). But while a further research will focus on more detailed indicators to qualify individual positions (in particular, centrality measures, as detailed by Koschützki et al. 2005 or Newman 2010), this first overview still shows that some groups of disciplines form very obvious clusters. This is the case of physics (right), medical sciences (bottom left) or earth sciences (top right). Others are sparsely connected or dispersed within other communities, as is particularly the case for disciplines like economics/business studies or chemistry, which seem to be more engaged in interdisciplinary collaborations or projects that include a limited number of employees (large experimental science projects partly explain the density of these groups). We also assume the structure of the network to differ among disciplinary specificities and temporality (Bourdieu, 2004; Gingras, 2012; Heilbron & Gingras, 2015). Are most connected disciplines also the most prestigious ones?

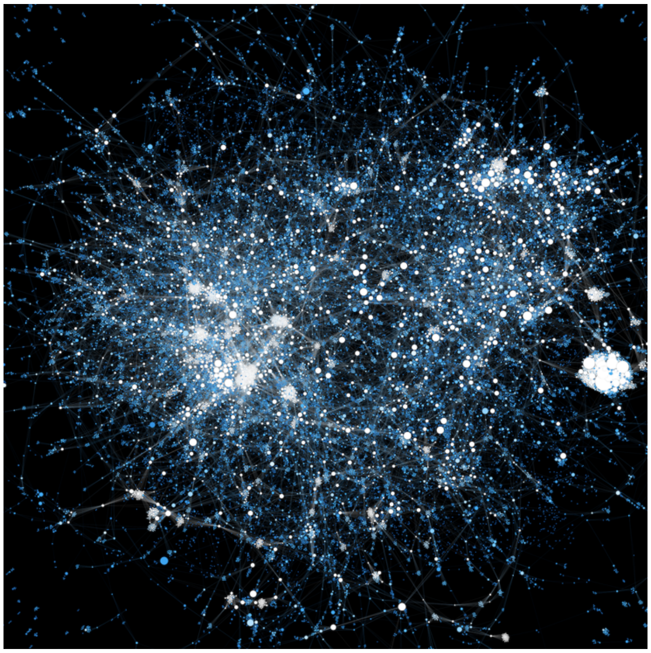


Figure 2. The core of the Swiss network of SNF scientific collaborations 2006-2016 (Grandjean 2016).

Perspectives

With the information contained in the list of projects, we see that it is also possible to assign individuals a disciplinary category extracted from the projects involving them. As it happens that a researcher is participating to projects labelled in different disciplines, this approach will lead to a reflexion on the measurement of interdisciplinarity within a comparative study between a selection of « open » and « closed » disciplines.

We will also see that it is possible to develop a multilevel analysis to compare the graph clustering to the many Swiss institutional and disciplinary « geographies », in order to historicize their development.

Mapping Prohibition: The Challenges of Digitally (Re)creating Historical Spaces

Hannah C. Griggs
hannahgriggs@gmail.com
Boston College

This poster will explore the challenges of creating digitized historical spaces as faced through the design and subsequent exhibition of the interactive, open-access map, *Prohibition Raids in New Orleans, 1919-1933*. During Prohibition, Federal agents (or 'Prohis') raided thousands of establishments throughout the city of New Orleans, arresting thousands more. Using data gleaned from *The Times-Picayune*, one of New Orleans's oldest newspapers, this map documents the proliferation of these raids through the 1920s into the early 1930s. Currently housed in The Museum of the American Cocktail in New Orleans, this exhibit is a collaboration between the Southern Food and Beverage Museum and my digital humanities project, Intemperance.org, an Omeka archive of cocktail culture in New Orleans. The goals and questions raised by this project will be presented in this poster. These include:

1. Summarizing the historical context of the project and considering the ways in which early 20th century New Orleans constructed and created spaces of leisure, as well as assessing the cultural importance of the bars, restaurants, residences, and other locales raided by Prohibition agents;
2. Exploring how to turn historical spaces into modern spaces in ways that are intuitive and accessible to diverse audiences and different publics;
3. Optimizing user experience for those viewers unfamiliar with the historical information presented and/or the digital platform(s) with which the exhibit was created;
4. Acknowledging the limitations of visualization and user experience of digitized historical spaces.

As a project rooted in Public History, the challenge of achieving the aforementioned goals lay in the digital platform. This poster will explain the exhibit's functionality goals and specific interactivity issues encountered during the exhibition's creation. This includes the ability to

- optionally read a description of the exhibit and summarized history;
- view details of each individual Prohibition raid;
- organize and view raids by year;
- filter by type of establishment raided;
- zoom in, zoom out, and view the page in full.

In order to achieve these goals, the exhibit was initially built in Neatline using the Starter Theme from Scholars' Lab, which works capably on any modern browser. Neatline advantageously allows users to explore the breadth of the historical archive by connecting Neatline points to Omeka items. When displayed on the Museum's multitouch tablet, however, issues of mobile compatibility and long loading times resulted in poor user experience. Ultimately, we had to modify the exhibition by changing the visualization platform to Tableau Public. Tableau not only circumvents the mobile compatibility issues found with Neatline, but it allows users to interactively filter historical data by year and raid type, a feature not currently possible in Neatline.

While both Neatline and Tableau are capable platforms to achieve these goals, each has its own advantages. With these specific goals in mind, this poster will furthermore compare and contrast the advantages and limitations of Neatline and Tableau Public as both mobile and desktop interfaces. This poster will furthermore highlight the exhibition's successful outcomes as well as the limitations of its interactivity. Much could be learned about visualization from the challenges faced in the creation of this exhibition.

Comparing Correspondences: Robert Southey and Willa Cather

Gabriel Hankins
ghankin@clemson.edu
Clemson University

Frank Elavsky
frankelavsky@gmail.com
Northwestern University

This poster presents recent work in the visualization of literary correspondence from the TCLLP (Twentieth-Century Literary Letters Project), focusing on the comparative visualization and analysis of digital editions of letters.

How might we begin to compare the collected correspondences of letter-writers? As archival letters are increasingly digitized, either by libraries and special collection or by the large-scale digital correspondence projects now emerging (e.g. the Electronic Enlightenment project, Early Modern Letters Online), we now have the opportunity to investigate authorial correspondence at larger scales. Beyond the usual use of correspondence to fill out the historical or biographical background of a writer, or to provide thematic and social points of linkage between writers, large collections of digital letters allow us to pose a range of new questions and problems: what does the social network of highly interconnected writers look like? What sorts of visual arguments and tools are appropriate to inquire into literary correspondence? How do the forces of class, gender, status, politics and profession manifest in a large correspondence? How do we read the gaps, cracks, elisions, and unseen limitations inherent both in our archival evidence and our digital tools? What are the large-scale shifts in epistolary culture over time, and how might we better understand those shifts in the correspondence of particular writers? To think through those questions more concretely, we visualized and compared the metadata of two well-curated digital correspondences from the early nineteenth and early twentieth century, the *Collected Letters of Robert Southey* and *A Calendar of the Letters of Willa Cather.*

We worked with the editors of these letters, and experts in TEI and RDF metadata transformation, to move the relationship implied by letter metadata – information about personal, political, geographical, and editorial networks of these very different writers – into visual representation. Our work forms part of an ongoing effort to bridge the gap between the Textual Encoding Initiative community, open linked data initiatives, and wider scholarly, library, and public audiences, as well as investigating important issues in data visualization as a hermeneutic activity.

This poster session presents our refined visualizations of these materials, and includes procedural as well as literary-critical findings: we will outline the process by which we move from TEI or less structured correspondence data into graph and network visualization; note some of interpretive and archival issues involved in that transformation; and display the refined visualizations comparing the political, personal, and literary networks of Southey and Cather. We hope to have an interactive conversation with attendees interested in network visualization, literary correspondence, Romanticism and modernism, open linked data, and the new TEI standards for correspondence data. The TCLLP is an open group of scholars interested in new approaches to twentieth-century literary correspondence, across linguistic, national, and disciplinary divides: we look forward to a discussion with Francophone scholars in particular.

Initial visualizations will be available on the project blog.

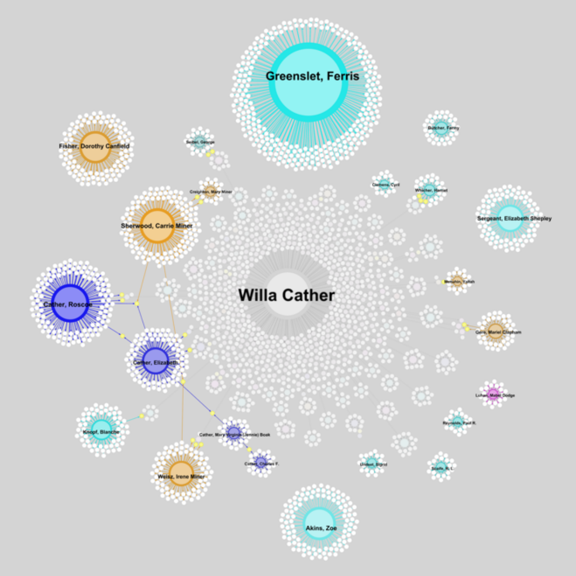


Figure 1: Willa Cather’s major correspondents

Bibliography

Twentieth Century Literary Letters Project (n.d.) . Project Blog. http://www.modmaps.net/tcllp/project-blog/

A Case Study of Automated Curation of Digital Archives

Fabiola Hanna
fhanna@ucsc.edu
UC Santa Cruz, United States of America

Access to digital archives has been well problematized in recent years. For example, should one have access to an archive by default or should one belong to a community in order to gain access to that knowledge,– such as with Mukurtu CMS which builds on knowledge heritage in indigenous groups (Christen 2007)? Another thread with regards to access and archives is the trend of dumping all the data, and claiming that because of this, the individual and/or the organization is somehow more transparent (as seen with various initiatives such as data.gov). But many have also shown that there is no raw data (Gitelman 2013). I propose to build on these two threads in order to argue that 1) paying attention to the medium and its parameters is important, and 2) that archives need to be sifted and curated in order for them to be properly accessible. I will illustrate both of these arguments with my project, *We Are History: A People’s History of Lebanon*, which is in its final stage of development and will have been released publicly before DH2017.

Digital Humanities (DH) projects are, in varying degrees, led by the desire to engage with a wider public. Some often include actions such as inviting participants to share their stories, images, audio clips, drawings, and videos. Many DH projects place these contributions in an audio or video database displayed in full on a webpage, not unlike oral history transcripts ending up in a dusty closet. To build on one of the most successful digital storytelling projects, it is useful to examine the Storycorps team, who have been collecting ordinary oral histories in video form and archiving them in full as a record of American history using booths and a mobile application. The Storycorps team knows very well that if it did not curate and edit together shorter versions, then few users would listen to the longer interviews in full, let alone several at a time.

Francis X Blouin and William G. Rosenberg trace the intersection of history and archives and found that Ranke, during the Enlightenment, conceptualized history as a scientific endeavor in that truth could be extracted from archives through rigorous methodologies. This led to the idea that documents could "speak for themselves" (Blouin, 24), as if simply making documents available, without providing context of any sort, would reveal their inner truth. This is one of many cases where the reading of documents is taken for granted. It also ignores the effect that archivists have on the collecting, saving, and indexing of documents. Influential archivists such as Terry Cooke, Richard Brown and Brian Brothman have brought about new attitudes to repositories with an acknowledgment of the effect that archivists have on documents (as quoted by Cox, 33). This relatively recent push in archival theory, therefore, points to the flaws in the claim that documents on their own can represent themselves: that would be ignoring all the various power relationships at play, as well as the medium itself in which the data is codified.

This is more directly seen in the tagging of videos and their categorization without additional interpretational work, such as in the Oral History Metadata Synchronizer (OHMS) tool developed at the Louie B. Nunn Center for Oral History University of Kentucky Libraries. This *will to not "add"* to these stories seems to come from the premise that these testimonies should "speak for themselves"; that no added interpretation is needed, even that any added interpretation distracts from the directness of the stories. But this often also means the medium and its effects on these stories are not carefully examined.

In pursuit of generating communal dialogue in the context of inability to have conversations about our contested history in Lebanon, I set out to build an Artificial Agent that would sift through an oral history video archive of testimonies of daily life with the task of figuring out common threads, sometimes confirming and sometimes contesting each other, and automatically editing many different versions of possible histories. This automatic montage machine addresses two problems in the Lebanese context: first, it circumvents the tiring accusation of being biased since a machine is now the moderator (presenting a multiplicity of stories might be the closest one can get to strategic objectivity) and second, it opens up the possibility of conversation by weaving various and often opposing perspectives in order to start imagining what our histories could look like. The project, which would reside online as well as in booths in public spaces across Lebanon, invites people to listen to an automated montage of oral histories and to then share their own stories and memories. Each newly contributed story is added to the archive, analyzed using new developments in computational corpus-based linguistics, automatic story generation, and social computing, and tagged with its transcript, which enables the interface to incorporate newly added video interviews into the pool concerning the event discussed

Mapping Pliny's Social Network: A Case Study in Digital Prosopography

Benjamin Hicks
bhicks@princeton.edu
Princeton University, United States of America

Introduction

This proposed short paper examines the progress and some preliminary observations of a practical attempt to apply digital humanities methods to the letters of a second century Roman aristocrat Pliny the Younger. The preliminary results of this study are available at the Pliny Project site. Its aim is to present a case study of how an initial research idea can be expanded to connect with larger digital humanities work in a particular field.

These letters, written in the early second century CE, are a treasure trove of social and literary information about the Roman elite during the period in which Roman territorial control reached its apex. As one of the most extensive collections of letters from the ancient world, as well as one of the most thoroughly explored, they are a rich data set on which to draw. They are paralleled by only a handful of similar letter sets from the Roman world, such as the letters of Marcus Tulilus Cicero or the four century orator Libanius.

In a broader context, Classics has been a frequent and early adopter of digital humanities methods (see *Online Coins of the Roman Empire*, the literary comparative tool *Tesserae*, and the venerable *Perseus Digital Library).* The field has also produced some initial attempts to bridge various "people indices" into a standard prosopography (i.e. the Standards for Networking Ancient Prosopographies, hereafter SNAP). Pliny's correspondents have been integrated into such resources, but they are either often limited to major university research collections or so unwieldy as to make consultation difficult. Examples of this tension include the standard tool for the prosopography of the Roman Empire, the *Prosopographia Imperii Romani*, 2nd edition (de Gruyter: 1933-2015) (= *PIR*^{2}), which has reached a massive eight parts with numerous fascicles. Only the index is widely available online. Likewise the most recent work on Pliny's names, Anthony Birley's *Onomasticon to the Younger Pliny (*Birley, 2012*)*, exists as a traditional monograph, albeit with a searchable PDF.

Such resources, though tremendously important to scholars specializing in the field, often constrain access for the broader academic community. Moreover, the rich information they provide is not structured in a way for easy search and access.

The research project on which this paper draws developed from a November 2015 – January 2016 affiliated fellowship at the American Academy in Rome, which gave me access to the prosopographic material in a single, well-organized location. Its primary objective was, and is, to create a comprehensive resource for Pliny's social network with an emphasis on the social class of his correspondents. My initial inquiries centered on compiling a list of Pliny's correspondents and attempting to identify them as best possible. The conventions of Roman naming, which resulted in many similar names within family groups, renders this difficult. The use of only a single name (compared to the somewhat standard use of two names) in one of the surviving manuscripts of the letters further complicates the task. Even if a family and identity of an individual is known, his or her social standing may not be clear. The Roman distinctions between a common citizen, the middling administrators of the equestrian class, and the upper rungs of the senatorial class were very sharp to them—so sharp they often saw no need to clarify who was of what class for posterity.

This made data modeling and cleaning a significant challenge, for which I employed exploratory tools such as SocNetV and more recently Cytoscape, for exploratory visualization. (Note: this issue of authorial ambiguity is not new to DH and letters. It has been frequently confronted by projects such as Stanford's *Mapping the Republic of Letters*). Some preliminary results are available through Pliny Project (see above), but they have been revealing both in terms of confirming known associations and providing new clarity into the possible editorial methods Pliny used in selecting his letters, which were curated for publication within his lifetime.

In order to construct a data set for such an analysis, I attempted to model a degree of closeness of connection by assigning a weight based on the number of times Pliny either mentioned someone in a letter or wrote to them to a reciprocal connection. This was saved in GraphML format to construct a diagram of centrality with shading of points to indicate the social class of Pliny's correspondents.

My talk will, in addition to discussing the above methodology in greater detail, center on two examples of preliminary results from this research, and then turn to future plans for the project. First, the set of social acquaintances that have often been associated with as what is informally called "Pliny Country"—near his home near modern-day Como, Italy—and the set associated with the city of Tifernum, both appear clearly in the social network map as a set of closer intimates, largely from the same equestrian class of which Pliny's family originated (the original formulation comes from the work of the eminent historian Ronald Syme, see Syme, 1991, for his collected works and the exploration of some of Pliny's connections of Tifernum in Champlin, 2001).

This gives preliminary confirmation that the methodology of simple weighting based on mentions as some approximation of closeness can be used in analyzing his social network.

A secondary observation is a series of correspondents to whom Pliny writes roughly two to three letters in the second to outer circle of his acquaintances. Some of these individuals are men who had held the consulship, the highest office to which a Roman not of the imperial family could aspire and all were of the senatorial class. Pliny rose to that same class from middling origins during his career, thanks to the patronage of his uncle and adoptive father. That there is a cluster of these letters with a remarkably similar number speaks to an editorial hand at work in their selection. While at this point identifying a motivation is primarily speculative, at the least we can say that it reveals a trend not previously identified and demonstrates an editorial concern for cultivating Pliny's prestige by association.

In addition to the specific application of this data to my own research, the longer term goals of this project are to provide this same dataset, edited and curated, to the broader scholarly community. I have currently published a simple database interface that allows users to search for Pliny's correspondents and note which letters are written to them. While this may seem on the surface a straightforward question, by integrating current scholarship and attempting to identify correspondents fully, it presents new and easier access for scholars, regardless of institutional affiliation.

Nevertheless, the initial search functions, which allow a name search and a tentative search by social class, are not sufficient to realize the goals of the project. My current development work is focused on transitioning the database to using Django's web application functionality and object database modeling to allow for the relationships noted in my social network analysis to be available and searchable. This transition to a standard platform will also lead to a web application that can be cloned from a tool such as GitHub and used by DH scholars to build or innovate using my dataset. It also acknowledges the need to connect this new structuring of the prosopographic corpus for Pliny to the broader initiatives to create people indices by including links to *PIR*^{2} search masks and SNAP.

Such an approach takes the traditional field of Plinian prosopography and attempts to open it to a wider scholarly audience. It also emphasizes the importance of exploratory visualization techniques in examining datasets for novel connections. This discussion will offer the audience an opportunity to consider how a project focusing on a specific area of research can connect with larger scholarly endeavors in DH.

Bibliography

Birley, A. (2012). *Onomasticon to the Younger Pliny.* de Gruyter.

Champlin, E. (2001) "Pliny's Other Country," in *Aspects of Friendship in the Greco-Roman World*, ed. Michael Peachin (Journal of Roman Archaeology), 121-128

Syme, R. (1991) *The Roman Papers*, vol. 2, Oxford, 694-698.

POSTagging and Semantic Dictionary Creation for Hittite Cuneiform

Timo Homburg
timo.homburg@hs-mainz.de
Mainz University of Applied Sciences, Germany

Presentation Topic and State of the Art

On our poster, we want to present ongoing work to create an automatic natural language processing tool for Hittite cuneiform. Hittite cuneiform texts are, to this day, manually transcribed by the respective experts and then published in a transliteration format (commonly ATF). Pictures of the original cuneiform tablet may be provided and more rarely cuneiform representations in Unicode are present. Due to recent advancements in the field (such as Cuneify) an automatic translation of many Hittite cuneiform transliterations to their respective cuneiform representation is possible.

Research Contributions

We build upon this work by creating tools that aim to automatically translate Hittite cuneiform texts to English from either a Unicode cuneiform representation or their transliteration representation.

POSTagger

We have created a morphological analyzer to detect nouns, verbs, several kinds of pronouns, their respective declinations and appendices as well as structural particles. On a sample set of annotated Hittite texts from different epochs in cuneiform and transliteration representation, we have evaluated the morphological analyzer, its advantages, problems and possible solutions and intend to present the results as well as some POSTagging examples in section one of our poster.

Dictionary Creation

Dictionaries for Hittite cuneiform exist in often non-machine readable formats and without a connection to Semantic Web concepts. We intend to change this situation by parsing digitally available nonsemantic dictionaries and using matching algorithms to find concepts of the English translations of such dictionaries in the Semantic Web e.g. DBPedia or Wikidata. Dictionaries of this kind are stored using the Lexical Model for Ontologies (Lemon). In addition to freely available dictionaries we intend to use expert resources developed by the academy of sciences in Mainz/Germany to verify and extend our generated dictionaries. We intend to present the dictionary creation process, statistics about the content of generated dictionaries and their impact in section two of our poster.

Machine Translation

Using the newly created dictionaries as well as the POSTagging information we intend to test several automated machine translation approaches of which we will outline the process and possible approaches in poster section three.

Contributions for the Communities

With our approaches, we intend to contribute to the archaeological community in Germany by analysing Hittite cuneiform tablets. Together with work from the University of Heidelberg on image recognition of cuneiform tablets, we want to focus on creating a natural language processing pipeline from scanning cuneiform tablets to an available translation in English.

Spatial-temporal Variation based Innovation History Visualization: A Case Study of the Liquid Crystal Institute at Kent State University

Tao Hu
taohu07@hotmail.com
Kent State University, United States of America

Marcia Zeng
mzeng@kent.edu
Kent State University, United States of America

Yin Zhang
author.email@domain.com
Kent State University, United States of America

Xinyue Ye
xinyue.ye@gmail.com
Kent State University, United States of America

Hongshan Li
hli@kent.edu
Kent State University, United States of America

Innovation began taking root as a term associated with science and industry in the nineteenth century, matching the forward march of the Industrial Revolution, although the language of that period focused more strongly on the invention, particularly technical invention (Green, 2013). Nowadays, innovation is defined simply as a "new idea, device, or method" (Wikipedia). In the Big Data era, innovation history research not only shows the raw data but also demonstrates and reveals the deep relationships of data and how the innovation was generated.

This poster mainly describes an innovation history analysis system, integrating and visualizing multi-source and isomerism data in a static and dynamic representation manner, considering spatial-temporal factors. A case study of the Liquid Crystal Institute (LCI) at Kent State University is presented. Liquid Crystal Institute (LCI) was founded in 1965 by Glenn H. Brown, a chemistry professor at Kent State University. The birthplace of liquid crystal displays (LCD), the LCI is the world's first research center focused on the basic and applied science of liquid crystals. The dramatic rise of the liquid crystal display (LCD) industry through the subsequent 40 years has fundamentally changed our modern life.

As shown in Figure 1, to research the innovation history of LCI, there are varieties of data sources in the forms of audio, video, digitalized images, text from the website, annual reports from 1965-2013, interviews with key researchers in LCI, booklets from LCI's 50 year anniversary that covers significant scientists and important events, as well as biennial International Liquid Crystal Conference (ILCC) materials about the largest academic meeting in the field of liquid crystals that was started by the LCI founding director Glenn H. Brown in 1965. At the data processing level, data searching rules related to the publications, grants, and patents are appointed. For text, text mining tools, such as Open CALAIS and Cogito Intelligence API, are used to extract people, locations, and event information from annual reports, booklets etc. After the initial processing of raw data, data are then imported into our system, including publications records, grants, patents, inventions, researchers (such as research staff, post-doctoral fellows, visiting scholars, etc.), special events, spin-offs, and conferences.

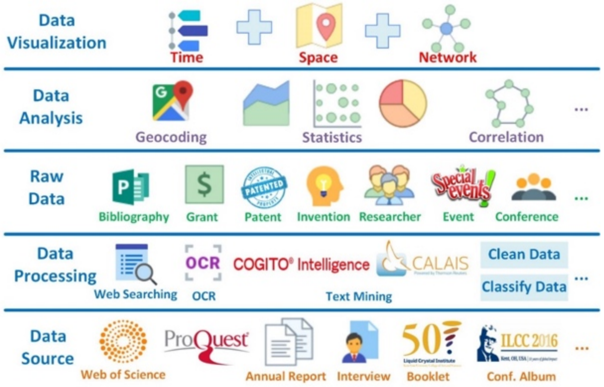
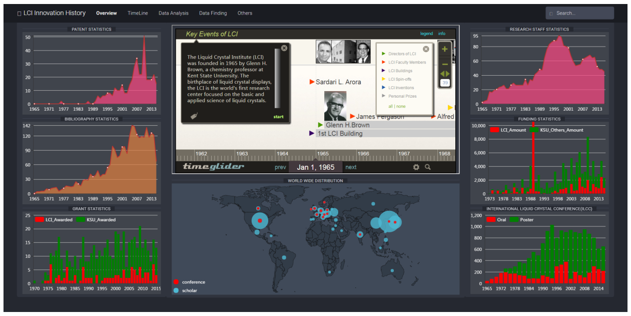


Figure 1. Structure of LCI innovation system

As argued by Robert E. Williams (Williams, 1987), location can be the critical link to integrating data from various sources and with various attributes of a place. Using locations as the key link, any location-related description, from the formal coordinates of places to the informal abstraction of places, can be understood by using GIS and various types of Big Data. Thus, the system geocoded the data which has location information. Besides traditional data statistics methods, with geo-located data, spatial statistics algorithms can also be imported to our system to find the spatial correlation of data.

Data visualizations are used to demonstrate the worldwide impact of LCI related scientists, technologies, and events. Unlike other tools, such as the "Historical Data Exploration Tool" (Škvrňák and Mertel, 2016), this system will statically visualize time, space and network variables, whilst also considering the dynamic network changes in each period. Furthermore, spatial-temporal analysis algorithms will be imported to display the spatial and temporal pattern of LCI development. Figure 2 shows visualization results based on the gathered and processed data. It is developed by modern web technologies (html5 canvas, javascript, d3 – see Bostock et al, 2011; timeglider). In the center top, there's a timeline integrating important faculty members, inventions, personal prizes, spin-offs, and other great events since 1965.

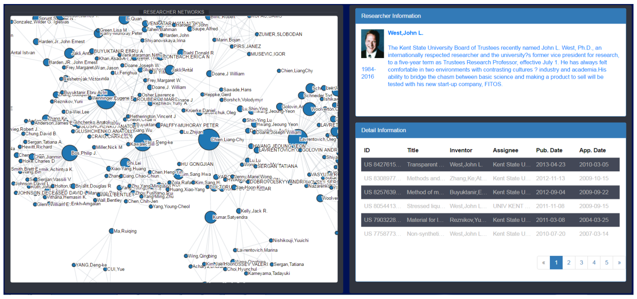


The Liquid Crystal Institute (LCI) was founded in 1965 by Glenn H. Brown, a chemistry professor at Kent State University. The birthplace of liquid crystal displays, the LCI is the world's first research center focused on the basic and applied science of liquid crystals.

Figure 2. Main interface of innovation history research system

Scientific collaboration is a complex social phenomenon in a search that has been systematically studied since the 1960s (Noldus and Van Mieghem, 2015). One of the goals of developing this system is to find out how the LCI can be so successful in liquid crystal field research and development.

From the view of research collaborations, it may demonstrate the answers. Thus, the system processes the data related to papers, patents and grants, and visualize the collaboration networks shown in Fig. 3. On the left of the figure, it demonstrates the networks, in which each node indicates an author (inventor or grant recipient) and each link indicates the collaboration. On the right, it presents the detailed information of researcher when clicking on the node.



West,John L.

1984-2016

The Kent State University Board of Trustees recently named John L. West, Ph.D., an internationally respected researcher and the university?s former vice president for research, to a five-year term as Trustees Research Professor, effective July 1. He has always felt comfortable in two environments with contrasting cultures ? industry and academia.His ability to bridge the chasm between basic science and making a product to sell will be tested with his new start-up company, FITOS.

ID	Title	Inventor	Assignee	Pub. Date	App. Date
US 8427615...	Transparent ...	West,John L.	Kent State U...	2013-04-23	2010-03-05
US 8308977...	Methods and...	Zhang,Ke;At...	Kent State U...	2012-11-13	2009-10-15
US 8257639...	Method of m...	Buyuktanir,E...	Kent State U...	2012-09-04	2009-09-22
US 8054413...	Stressed liqu...	West,John L....	UNIV KENT ...	2011-11-08	2009-09-15
US 7903228...	Material for l...	Reznikov,Yu...	Kent State U...	2011-03-08	2004-03-25
US 7758773...	Non-syntheti...	West,John L....	Kent State U...	2010-07-20	2007-03-14

Figure 3. Collaboration visualization of patent awardees

This system can be used not only to further relevant historical research but also to serve as a prototype to demonstrate the potential for linking different visualization techniques to provide functions facilitating historical data exploration. Furthermore, it provides collaboration network analysis environment to dig out new findings in innovation history.

Bibliography

Andrienko, N. and Andrienko G. (2006). *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach.* Berlin ; New York: Springer.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D3: Data-Driven Documents, *IEEE Transactions on Visualization and Computer Graphics*, 17(12): 2301-09.

COGITO Intelligence API (n.d.). http://www.intelligenceapi.com/. November, 2016.

Green, E. (2013). Innovation: The History of a Buzzword. *The Atlantic.* 20 June 2013. http://www.theatlantic.com/business/archive/2013/06/innovation-the-history-of-a-buzzword/277067/. November, 2016.

Noldus, R. and Van Mieghem, P. (2015) Assortativity in complex networks. *Journal of Complex Networks*, 2015.

Škvrňák, J., and Mertel, A. (2016). Linking Graph with Map for the Purpose of Historical Research. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 887-888.

Von Lünen, A., Travis, C. (eds). (2013). *History and GIS: Epistemologies, Considerations, and Reflection.* Springer, Dordrecht.

Wikipedia Contributors (n.d.) Innovation. https://en.wikipedia.org/wiki/Innovation. November, 2016.

Williams, R. E. (1987). Selling a geographical information system to government policy makers. *Annual Conference of the Urban and Regional Information Systems Association.*

Add Slander and Stir: Reprinting of Anti-Labor Stories in Early-Twentieth-Century American Newspapers

Vilja Hulden
vilja.hulden@colorado.edu
University of Colorado-Boulder
United States of America

In the opening years of the twentieth century, the National Association of Manufacturers (NAM), an organization of businessmen vehemently opposed to labor unions, spearheaded a multifrontal and multi-industry campaign aiming to stop the rapid growth of labor unions. The NAM itself envisioned publicity and propaganda – or a "campaign of education," as the NAM liked to call it – as a key aspect of this campaign. Newspapers, in turn, were to play a crucial role in the campaign. As the head of the publicity agency the NAM hired said, the "infantry" of the whole campaign consisted of "the constant repetition of small paragraphs or short articles throughout the newspapers of the country.”

The publicity bureau promised that getting this “infantry” into the field was going to be smooth sailing, claiming that "it is easily possible to bring together the vast majority of papers all over the country in a kind of syndicate form." With this, the bureau referred to syndication of the type now familiar to us, but also something specific to the early 20th century: boilerplate and readyprint. Because typesetting was still very expensive, large numbers of small country newspapers got part of their paper either preprinted (readyprint) or as boilerplate (stories that were already typeset and could be directly printed).

Archival sources show that the NAM, besides hiring a publicity bureau to badmouth labor unions, also created "newsworthy" events that it hoped to get into the press without direct payment: these included, for example, staged anti-union parades and Astroturf anti-union worker organizations. What we do not know with much certainty, however, is to what extent these efforts were successful.

This poster reports on a preliminary investigation into the extent to which stories reproducing the language or viewpoint of anti-labor organized employers – a language that is quite recognizable if one is familiar with it – were reprinted in early-twentieth-century American newspapers. It draws on the Chronicling America database of newspapers and makes use of a reprint detection algorithm developed for the Viral Texts project. Eventually the aim is to create as comprehensive an account of such reprinted texts as allowed by the geographical and other limitations of the Chronicling America collection.

Sources and methodology

The dataset used in this investigation consists of the articles (or rather, pages) published between 1897 through 1908 that matched the query (worker OR workers OR workmen OR workman) AND (union) in the Chronicling America collection. The time period selected reflects the years surrounding the launching and high point of the employer open shop campaign, including the NAM's hiring of the publicity bureau noted above.

The identification of reprints is performed using the passim tool developed by David Smith for the Viral Texts project with the default settings (which were developed on Chronicling America material). The results are formatted into an HTML form for easier manual checking; as the algorithm inevitably identifies irrelevant reprints, the most laborious stage is the culling of the set of identified repeated passages.

The early-twentieth-century employer anti-union campaign had a recognizable vocabulary and set of themes, and it is the author's expertise, acquired developed through research on these employer campaigns, that is the main "method" for recognizing stories potentially put out by employers.

Results

As of current writing, I have processed 1906, 1907, and, partly, 1908. For these years I have run the passim algorithm and manually examined the resulting identified reprinted clusters, which has turned out to be even more laborious than anticipated.

On the whole, I expected to see more stories clearly planted by NAM or its allies in readyprint or boilerplate. There are a number of candidates, but these are both less conclusive and have fewer reprints than I anticipated. There are, however, a few very widely reprinted clearly paid-for advertisements excoriating unions.

In some ways, the partly "negative" results in themselves are rather interesting (they seem not to be at least entirely due to any failings of the algorithm – widely printed items like speeches by the President are also prominent among the reprints). They seem to indicate that despite publicity bureaus' interest in claiming that material could be easily and discreetly planted, the task of shaping public opinion was more challenging that the NAM perhaps anticipated. The very fact that readyprint and boilerplate had to be palatable to a broad cross-section of papers, in an environment where there still were many labor or pro-worker papers, meant that "syndicated" material needed to be bland and apolitical to sell. Perhaps the democratic role of the press was better protected than expected?

Bibliography

Baldasty, G. J. (1999). *E.W. Scripps and the Business of Newspapers*. Urbana: University of Illinois Press.

Cordell, R. and Smith, D. A. Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines. http://viraltexts.org/.

Guarneri, J. A. (2012). Making Metropolitans: Newspapers and the Urbanization of Americans 1880-1930. Ph.D. thesis, Yale University.

Harter, E. C. and Harter, D. (1991). *Boilerplating America: The Hidden Newspaper.* Lanham, MD: University Press of America.

Schramm, W. and Ludwig, M. (1951). "The weekly newspaper and its readers." *Journalism Bulletin*, 28(3): 301–14.

A Seat At "La Tawola"

Michael Iannozzi
miannozz@uwo.ca
Western University, Canada

Introduction

In Italian, il tavolo means 'the table', but la tavola, the feminine form, is 'the sense of family felt around the dinner table'. My grandparents taught me the meaning of la tavola first-hand; they emigrated from Italy in 1959. Through my research, I document the Italian dialect of immigrants in Sarnia, Ontario, Canada. Through community collaboration, we work to preserve their stories.

What?

Heritage languages are those brought by immigrants to a new place, and now spoken as a minority language by that community (Fishman, 2001). There are two parts of this research: to record the histories of this community in their native Italian dialect; and second, to create a permanent and public digital archive to preserve this Heritage Language.
The research question is: What is the best approach to create a community-university partnership that will benefit both the heritage and academic communities?

Why?

From 1950-1969, at Halifax's Pier 21, 20,000-30,000 Italians arrived each year in search of a better life (pier21.ca). Today, 1.4 million Canadians identify with an Italian heritage (StatsCan, 2013). Almost all Canadians have a 'hyphenated' identity: Italian-, Dutch-, French-, etc. By collaborating with the university and heritage community to build a digital archive, I will establish an interdisciplinary means to preserve Heritage Languages and the histories of Heritage Communities throughout Canada.

Where?

I have signed a two-year partnership with the Archives of Sarnia and Lambton County to create an online archive. Through this project, I digitized and uploaded hundreds of recordings, videos, images, and documents recently collected by myself, and by Caroline Di Cocco since the 1980s. Caroline is a local community historian who has contributed what she collected over the past 40 years. We will formally present the archive to Sarnia's Italian community in June, 2017. This serves as a starting point showing the project's feasibility.

I have been given permission by Western University's Libraries to host this archive, permanently, at the University. The *Western Archive of Dialects and Languages* (WADL) is available for other language projects, and my Italian research is the initial project to be hosted there (www.ir.lib.uwo.ca/wadl). Western University is only an hour from Sarnia, which allows me to remain an active member of Sarnia's Italian community, and to keep the digital archive local.

How?

With the support of Western Libraries and Sarnia's Italian community, the archive is being created through my coordination of the community and academia. Through my work with Lambton County Archives: I code the website; use *Dublin Core* standards for metadata (Kunze et al., 1998); and digitize materials. I use Omeka software for the framework.

The archive is made up of many photographs, documents, and videos. I have collected and digitized just under 500 photographs so far. It is also made up of documents: digitized copies of boat tickets to cross to Canada, citizenship papers, letters, (delicious) recipes, and other ephemera of their Italian and Canadian experiences. I have just under 200 documents so far.

I am currently recording interviews with Ciociaria dialect speakers, using sociolinguistic techniques to collect life histories. I also have 47 interviews I've digitized that were conducted in the 1980s by Caroline di Cocco.

Conclusion

WADL ensures that the community directly benefits from sharing their stories, materials, and time with the researcher. The information they share is returned in a permanent, secure, and digital way. This ensures that future generations will be able to access their heritage and history.

The goal of this research is to show the importance of heritage—a significant part of the Canadian identity. My work will produce a collaborative partnership between Western University and Sarnia's Italian community. This framework can be applied to preserve and promote other communities across Canada.

In summary, the stories, culture, and language of this community cannot wait any longer to be recorded and digitized. There is an urgency as Sarnia's original Italian community is reaching an advanced age. Now is the time to create an online presence: both for academics and for the public. There are research opportunities in this data, from a great number of disciplines. The community and broader public is interested in hearing these stories, seeing these photos, and cooking some of the recipes.

I study how Sarnia's Ciociaria pronounce "la tawola", and ensure the survival of what they mean when they say it.

Bibliography

Barberis, C. (1999). *Le campagne italiane dall'Ottocento a oggi.* Laterza.

Fishman, J. A. (2001). "300-plus years of heritage language education in the United States". In Peyton, J. K.; Ranard, D. A.; McGinnis, S. *Heritage Languages in America: Preserving a National Resource*. Washington, DC: Center for Applied Linguistics. pp. 81–98.

Kunze, J. A., Lagoze, C., & Weibel, S. L. (1998). *Dublin Core Metadata for Resource Discovery*. Resource.

Pier 21. (2015). Retrieved 11 April 2016, from https://www.pier21.ca/

Statistics Canada. (2013). *Immigration and ethnocultural diversity in Canada: National Household Survey*, 2011. Ottawa: Statistics Canada = Statistique Canada.

Points, Lines, Polygons, and Pixels: A Framework for Teaching & Learning Humanities through Visualization

Hannah Jacobs
hannah.jacobs@duke.edu
Duke University, United States of America

Introduction

Points, lines, polygons, and pixels, the primary elements of digital visualization, can be arranged and rearranged to present infinite interpretations of space, time, objects, or patterns. Visualization offers a significant opportunity for humanities students to develop digital visual literacies through exploratory making, analysis, and storytelling. Yet incorporating digital visualization components into teaching depends upon instructors' course content and access to expertise, time, and tools. A scalable, replicable approach is needed to support visualization pedagogies that can be implemented in diverse educational environments. In this poster, I propose a framework for designing and implementing visualization projects within existing courses, drawing on my instructional experience in the Wired! Lab at Duke University.

Background

The rising prevalence of visual media in global society points to an increasing need to cultivate digital visual literacies in humanities classrooms. Susan Brown has noted this "shift from textuality to visuality," calling for more "active engagement with new technologies rather than passive consumption." Through this engagement, students may learn to create scholarly visualizations and can become thoughtful critics of the visual media with which they interact in everyday life.

My role with the Wired! Lab involves working with instructors to design and implement such visualization projects. These projects have involved a variety of methods, including architectural modeling, mapping, digital exhibit curation, and information visualization. (see the list of recent student projects) Through these collaborations, our instructional teams have recognized a need for both scalability and replicability in assignment design and implementation. Accordingly, I have begun to approach to project design with an eye toward creating instructional templates that can be applied to multiple topics and course levels. This strategy has been echoed by Aaron Mauro, who calls for an emphasis in undergraduate digital humanities on "iterability, openness, and extensibility" (Mauro, 2016). Assignments must be made flexible, or responsive, to varying constraints in instructors' and students' access to resources both in and beyond the Wired! Lab. In response to this need for project plans that can be adjusted to a variety of contexts, I am developing a framework for iterative course project design.

Framework

This proposed framework presents a decision-making process that enables instructors to shape project ideas into assignments suitable for their multiple contexts. The framework seeks to match digital methods and resources with pedagogical goals, course content, and visualization concepts by guiding instructors through an iterative planning process:

Instructors are first asked to devise a research question, based on their course goals and content, that will drive the assignment. Next they develop a list of possible source materials and ascertain whether these materials are accessible to students. If materials are not available, instructors revise their question around available sources. These sources are then examined with the aim of identifying types of information: is it quantitative, qualitative, or both? How might this information be further classified—spatially, temporally, statistically, categorically? Of these data types, which provide evidence needed to answer the research question? (If this process reveals that the data are not useful, instructors should return to their question and list of sources to reevaluate.)

Based on the process so far, instructors consider first how the visualization will be used: will it be a tool for exploring information and analysis? Or will its present an interpretive visual narrative? Second, will the visualization(s) be static or dynamic? Third, instructors consider which visualization type(s) might provide the best mode(s) of representation: scientific, informational, conceptual, spatial, temporal? Within these types, instructors begin to identify specific methods: mapping or 3D modeling, data visualization or visual narrative, for example. Then another moment of reflection: does the chosen method fit the research question, course goals, and visualization purpose? If not, instructors may reconsider their method in conjunction with these decisions.

Finally, instructors begin exploring possible tools as they examine their design thus far in comparison to both students' expected skill level and institutional resources: access to expertise (theirs or someone else's), additional time before and during the course, and tools (hardware, software, internet).

Resulting visualization projects may require few or many interventions in the course structure and content. They may rely on a combination of visualization techniques. They may be applied in multiple course iterations. Overall, they should engage the research question while advancing digital visual literacies.

Conclusion

Visualization is a powerful tool for communication that can be investigated across humanities disciplines. It can be employed to develop students' critical abilities to represent humanities research in a multitude of ways. This proposed framework addresses the challenges instructors face when implementing visualization in classrooms: from matching content to method, to identifying appropriate tools, to designing for scalability across one or more course settings.

Bibliography

Association of College & Research Libraries (2011) Visual Literacy Competency Standards for Higher Education, http://www.ala.org/acrl/standards/visualliteracy.

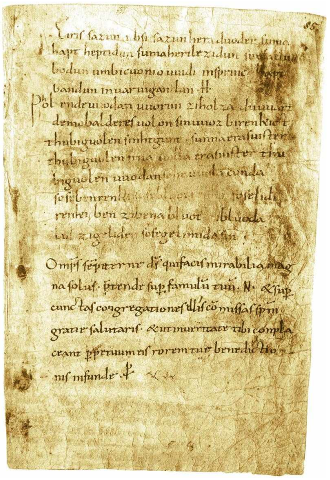
Brown, S. (2011). "Don't Mind the Gap: Evolving Digital Modes of Scholarly Production Across the Digital-Humanities Divide," in *Retooling the humanities: the culture of research in Canadian universities*, ed. Daniel Coleman. (Edmonton: University of Alberta Press), 210-211.

Mauro, A. (2016), "Digital liberal arts and project-based pedagogies," in *Doing Digital Humanities: Practice, Training, Research*, ed. Constance Crompton, Richard J. Lane, and Ray Siemens. (New York: Routledge), 379.

On the Impact of the Merseburg Incantations

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

While searching for medieval manuscripts suitable for the "Monumenta Germaniae Historica" (a collection of German historical sources) in 1841, Georg Waitz visited the library of the Merseburg Cathedral (Jankofsky, 2013). Rather coincidentally, he found a page with two magic spells in a theological composite manuscript: the Merseburg Incantations (see Fig. 1). Realizing the importance of his discovery, he asked Jacob Grimm to analyze and evaluate the text. The first Merseburg Incantation is a *blessing of release* telling about "Idisen" (female dieties) freeing either themselves or captured warriors from chains. The second Merseburg Incantation is a *healing spell* to cure a dislocated horse foot. The spells in Old High German were written down in the 10th century, but their origin is still unclear, maybe several hundred years earlier. The first time publicly presented at the Royal Prussian Academy of Sciences in Berlin, Jacob Grimm denoted the value of the Merseburg Incantations by calling them a gem, and nothing from the most popular libraries would have a similar value (Grimm, 1842).



Eiris ſazun idiſi ſazun hera duoder ſuma
hapt heptidun ſumaherilezidun ſumaclu
bodun umbicuonio uuidi inſprinc hapt
bandun inuaruigandun . H.
Phol endeuuodan uuorun zi holza du uuart
demo balderes uolon ſin uuoz birenkict
thu biguolen ſinhtgunt . ſunna era ſuiſter
thu biguolen friia uolla era ſuiſter thu
biguolen uuodan ſo he uuola conda
ſoſe benrenki [illegible] ſoſe lidi
renki ben zi bena bluot [illegible] bluoda
lid zi geliden ſoſe gelimida ſin .

Omps ſempiterne ds qui facis mirabilia mag
na ſolus . ptende ſup famulu tuu . N . & ſup
cunctas congregationes illis comiſſas ſpm
gratie ſalutaris . & ut in ueritate tibi compla
ceant ppetuum eis rorem tue benedictio
nis infunde .

Figure 1: Merseburg Incantations manuscript

A number of works interpret content, meaning and origin of the Merseburg Incantations (e.g., Schumacher, 2000, Beck, 2003, and Schmitt, 2011). In contrast to these rather profound analyses and interpretations of the Merseburg Incantations, our project aims at examining their global impact. We therefore design a platform that brings together various types of sources––tagged with citation year, source type, reference purpose––citing the Merseburg Incantations. Some examples are listed below.

Sources citing the first Merseburg Incantation:

- "Die Südharzreise", Frank Fischer, 2010, travelog
- "Mara und der Feuerbringer, Band 03: Götterdämmerung", Tommy Krappweis, 2012, novel
- "Angelina Jolie - The Lightning Star", C. Duthel, 2012, biography
- "Seelenriss: Depression und Leistungsdruck", Ines Geipel, 2013, psychology
- "Das siebte Buch: Objektorientierung mit C++", Ernst-Erich Doberkat, 2013, programming
- "Die Schwarzen Musketiere - Das Buch der Nacht", Oliver Pötzsch, 2015, novel
- "Charlemagne", Johannes Fried and Peter Lewis, 2016, historiography

Sources citing the second Merseburg Incantation:

- "The Key to Music's Genetics: Why Music is Part of Being Human", Christian Lehmann, 2014, musicology (example for healing through singing)
- "Ring of the Nibelungs", German dubbing, 2004, TV Movie (entertainment)
- "Die Leute vom Domplatz", Episode 1, 1979, German TV Series (entertainment)
- "To Ride a White Horse", Pamela Ford, 2015, novel
- "Healing Symbols in Psychotherapy: A Ritual Approach", Erik D. Goodwyn, 2016, psychology
- "Harzer Pferdezucht im Spiegel der Geschichte", Bernd Sternal, 2015, horse breeding
- "Metallische Implantate in der Knochenchirurgie", Erich Frank and Herbert Zitter, 1971, medicine (bone surgery)
- "Norse Magical and Herbal Healing", Ben Waggoner, 2011, medicine (medieval medical text collection)

Sources citing both Merseburg Incantations:

- "Handbuch der germanischen Philologie", Friedrich Stroh, 1985, philology
- "Götter und Kulte der Germanen", Rudolf Simek, 2004, mythology
- A number of music songs, e.g., performed by In Extremo, Saltatio Mortis, Corvus Corax or Tibetréa, music (entertainment)
- Merseburg Incantations Geocache, since 2004, geocaching puzzle

Although this is only a tiny snapshot of sources referring to the Merseburg Incantations, some tendencies are already visible. Whereas both Merseburg Incantations are cited in a number of books on mythology and philology (only one of each category is listed above), there are also some differences dependent on the spell contents, e.g., historiography for the first spell, and medicine for the second. Next to these reasonable source types, there are also unexpected findings like programming for the first spell, and horse breeding for the second. Both Merseburg Incantations are also often used for entertainment purposes in the form of novels, in music songs, or in movies. Also an interesting finding is a Geocaching puzzle guiding to a Geocache that is hidden in Merseburg.

The project is designed as a two-stage student internship, where Masters students of the humanities and computer science collaboratively work together in order to gain experiences for future digital humanities projects. The first step is data acquisition, which is performed by humanities students. Although web search engines, e.g., platforms with digitized contents such as Google Books or Internet Archive for textual sources, are the entry points for data acquisition, data extraction and structuring is done manually to ensure high data quality and to capture the purpose of the reference as precisely as possible. For example, we extract detailed information about the context in which a spell is cited, e.g., in the biography about Angelina Jolie the first Merseburg Incantation is referenced in a movie description of Beowulf, in which Angelina Jolie partook as a supporting actress. The second step of our project is the development of an interactive visualization system to support the dynamic exploration of the data collection. Therefore, computer science students design several visual interfaces that summarize different metadata information (e.g., a timeline to visualize the temporal impact, or a tag cloud to illustrate source types). One of the major functions of the system is a comparative view on the different contexts in which the Merseburg Incantations are cited together and each of them individually. Although the data is collected in a time limited project, the system is designed as a web-based crowdsourcing platform, so that the database can be extended after the project. A specific feature of the system is genericity, which makes it applicable to other texts in question. For example, creating a collection of citations of the Trierer Zaubersprüche or the Hildebrandslied will be possible, also in the form of a comparative analysis.

An interpretive approach using our proposed system will be complicated due to the fact that the collection contains serious scholarly as well as artistical references, and it is not possible to draw a clear line between such groups. Also, we consider each citation as equally relevant, which might not evolve a convincing representation of impact. While it is furthermore not possible to collect “all” citations, supporting hypotheses generation after distant reading analyses is not the prior purpose of our system. Our aim is rather to establish a starting point for exploring the far-reaching significance of the Merseburg Incantations––one of the most important written samples of German language, even the oldest and only known preserved text about Germanic Paganism in Old High German. With our project, we want to fulfill the responsibility to sustainably highlight this uniqueness.

Bibliography

Beck, W. (2003). Die Merseburger Zaubersprüche (Vol. 16). *Reichert Verlag.*

Grimm, J. (1842). Über zwei entdeckte Gedichte aus der Zeit des deutschen Heidenthums. *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin. Aus dem Jahre 1842.*

Jankofsky, J. (2013). Merseburg: 1200 Jahre in 62 Porträts & Geschichten. *Mitteldeutscher Verlag*.

Schmitt, M. (2011). Althochdeutsche Zaubersprüche als Textzeugen einer Zeit des Übergangs zwischen germanischem Heidentum und sich etablierendem Christentum – Form und Inhalt frühmittelalterlicher Magiepraxis. *Magisterarbeit*

Schumacher, M. (2000). Geschichtenerzählzauber. Die 'Merseburger Zaubersprüche' und die Funktion der "historiola" im magischen Ritual. In *R. Zymner (Ed.), Erzählte Welt - Welt des Erzählens. Festschrift für Dietrich Weber (pp. 201-215). Köln: Chora.*

Making topic modeling easy: A programming library in Python

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Christof Schöch
hristof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de
University of Würzburg, Germany

Topic modeling, a method for the semantic analysis of large text collections, has been in the focus of interest in digital literary studies during the recent years. The method uses probabilistic procedures to generate probability distributions for words out of a collection of texts, sorting many single word distributions into distinct semantic groups called ‘topics’. These topics constitute groups of semantically related words, and the contribution of each topic to the composition of each text can be quantified mathematically (Blei 2012, Steyvers und Griffiths 2006). In digital literary studies, topic models can be interesting in themselves. For example their dynamic development either during the plot of single literary texts or over multiple texts in a stage of literary history can be analyzed (Jockers 2013, Blevins 2012, Rhody 2012, Schöch to appear), though comparing literary themes and the probabilistic concept of ‘topics’ described here is obviously not unproblematic. And topic models can also be interesting features for classifying or clustering texts (Blei 2012).

There are currently two state-of-the-art implementations of the relevant algorithms: ‘Mallet’ (McCallum 2002) and

'Gensim' (Rehurek 2010). But usually more is required than simply running a topic modeling algorithm (Fig. 1):

- Longer texts like novels need to be split into smaller parts (e.g. paragraphs, scenes, or a fixed amount of characters or words).
- NLP based preprocessing is necessary
- To achieve optimal results, texts must be reduced to content words, either by filtering out function words with stopword lists, or by using a part-of-speech tagger to exclude unwanted word classes.
- Similarly, lemmatization and elimination of proper names can be useful.
- After the topics have been generated, results are usually visualized based on the relevant metadata.
- Results need to be evaluated with regard to internal or external criteria rather than just being left to interpretation.

The aim of our work is to provide digital literary scholars with a consistent, extensive and well documented programming library that allows them to do the necessary preprocessing, to generate topic models relying on the existing implementations, and to visualize and evaluate results within a single convenient scripting environment. We want to empower users with little or no previous experience and programming skills to create custom workflows mostly using predefined functions within a familiar environment (see the current stage of development on Github). Hereby, we want to lower the access threshold to topic modeling as a method, facilitating researchers in spending time experimenting with topic modeling and understanding how it generates results, rather than spending it for acquiring advanced technical skills before being able to try topic modeling at all.

The library will be developed for the programming language 'Python' that is well suited for NLP and data analysis tasks, and popular among digital literary scholars already. In addition, development can be partially based on functions from, and experiences made with a previous Python-based implementation of a topic modeling workflow developed by Christof Schöch et al., that can be regarded as a proof of concept.

A convenient tool for NLP analysis during preprocessing does exist in the DARIAH-DKPro-Wrapper (DDW) that covers a wide range of NLP tasks for many different languages and generates annotations in a Python-Pandas compatible format easily usable within our library.

Evaluating the results of topic modeling is not a trivial task but has proven to be rather challenging (Wallach et al. 2009, Chang et al. 2009). In order to evaluate topics we will provide functions for intrinsic evaluations, for example perplexity, and external evaluations, for example path length in a resource like wordnet. We will also support task based evaluation where topics are used for a classification task and evaluated on the basis of the results.

For the visualization of the results we can build on the tmw library mentioned above and provide plots of topics over time (or other dimensions), distribution of topics over texts using heat maps and others.

Functions will be designed with consistent syntax and in a way that allows users to grasp easily what they do and why, so users can combine them into scripts to implement their own project ideas with minimal learning effort. The ability to customize workflows will be facilitated by a thorough tutorial describing all functions, outputs and potential combinations in detail (see the tmw tutorial as an example for what we are planning).

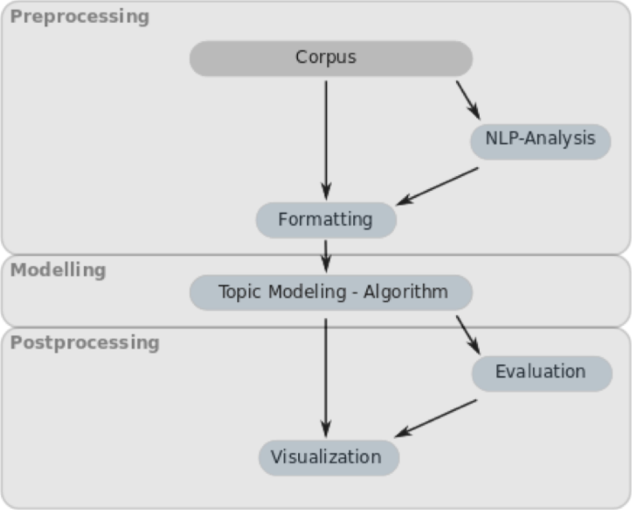


Figure 1: Topic modeling project workflow

Bibliography

Blei, D. M. (2012): „Probabilistic Topic Models". *Communication of the ACM* 55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

Blevins, C. (2010): „Topic Modeling Martha Ballard's Diary". *Historying*. http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.

Chang, J. (2009): Reading Tea Leaves: How Humans Interpret Topic Models. In: Y. Bengio et al.: Advances in Neural Information Processing Systems 22, 288--296.

Jockers, M. L. (2013). *Macroanalysis - Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.

McCallum, A. K. (2002): *MALLET : A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Rehurek, R., and Sojka, P. (2010): "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Rhody, L. M. (2012): „Topic Modeling and Figurative Language". *Journal of Digital Humanities* 2.1. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

Schöch, C. (to appear): „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama". *Digital Humanities Quarterly*. http://digitalhumanities.org/dhq. Preprint: https://zenodo.org/record/48356.

Steyvers, M., and Griffiths, T. (2006). „Probabilistic Topic Models“. In *Latent Semantic Analysis: A Road to Meaning*, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.

Wallach, H. M. (2009)et.al.: Evaluation Methods for Topic Models. In: Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.

Limiter l’impact des erreurs OCR sur les représentations distribuées de mots

Axel Jean-Caurant
axel.jean-caurant@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i) Université de La Rochelle, France

Cyrille Suire
cyrille.suire@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i) Université de La Rochelle, France

Vincent Courboulay
vincent.courboulay@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i) Université de La Rochelle, France

Jean-Christophe Burie
jean-christophe.burie@univ-lr.fr
Laboratoire Informatique, Image et Interaction (L3i)
Université de La Rochelle, France

Les chercheurs en Humanités numériques intéressés par l’analyse de grands corpus textuels utilisent de nombreuses méthodes et outils issus de domaines informatiques comme le traitement du langage naturel (Piotrowski, 2012) ou l’analyse de réseaux (Lemercier, 2005). Des méthodes récentes fondées sur les réseaux de neurones présentent également un intérêt majeur. Word2Vec est une méthode qui a grandement facilité l’utilisation de tels modèles (Mikolov, 2013). Les différentes optimisations apportées permettent, très simplement, d’entraîner un modèle sur de grandes quantités de données en utilisant un simple ordinateur de bureau. Le code source a été largement diffusé et a rendu cette méthode très populaire, notamment parmi les chercheurs en Humanités numériques. Hamilton a par exemple montré l’intérêt de ces modèles pour analyser l’évolution de certains mots du langage au cours du temps (Hamilton, 2016). Ces méthodes peuvent également être utilisées à d’autres fins. En effet, de nombreux corpus utiles aux Humanités numériques sont issus de processus de reconnaissance de caractères (OCR). Malheureusement, ces processus génèrent très souvent des erreurs, en particulier quand les documents analysés sont de mauvaise qualité (documents anciens ou mal numérisés par exemple). Ces erreurs touchent notamment les entités nommées comme les noms de lieux ou de personnes, particulièrement intéressants pour les chercheurs (Gefen, 2015). Ces erreurs ont un impact majeur sur l’accès à l’information car elles peuvent empêcher d’accéder à toutes les occurrences d’un mot d’intérêt.

Dans ce poster, nous présentons la méthode que nous avons développée pour étendre l’usage de Word2Vec à l’identification des erreurs OCR dérivées d’entités nommées. Après avoir entraîné un modèle sur un corpus donné, chaque mot est associé à un vecteur représentatif. Il devient alors possible de comparer les vecteurs pour extraire des relations morphologiques ou sémantiques entre les mots. On peut par exemple calculer la distance cosinus qui sépare deux mots dans l’espace vectoriel du modèle. Si, au sein du corpus, ces mots apparaissent dans des contextes similaires, la distance qui les sépare sera faible. Or, une entité nommée, bien que mal reconnue par le processus OCR, apparaît souvent dans le même contexte que l’entité originale. En combinant cette distance, qui agit sur les vecteurs, avec une distance d’édition sur les mots, nous pouvons identifier des mots proches sémantiquement et qui possèdent beaucoup de caractères en commun. Cette analyse produit ainsi une liste de termes qui ont toutes les chances d’être des entités mal reconnues par le processus de reconnaissance de caractères.

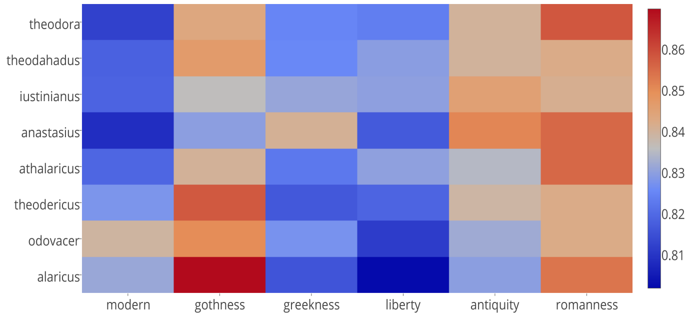


Figure 1: Expérience menée par Bjerva et. al., qui présente les similarités entre différents personnages et quelques grands concepts. Plus une cellule est rouge, plus la similarité est importante.

Une fois les erreurs identifiées, il est possible de s’intéresser à une entité nommée particulière. Sur la base des résultats précédents, nous proposons la construction d’un nouveau vecteur associant le vecteur de l’entité originale et les vecteurs représentatifs des erreurs. Ce nouveau vecteur est le résultat de la combinaison linéaire des vecteurs du mot original et des erreurs OCR. Pour modérer l’importance des vecteurs dans la combinaison, ces derniers sont pondérés selon le nombre d’occurrences du terme correspondant dans le corpus.

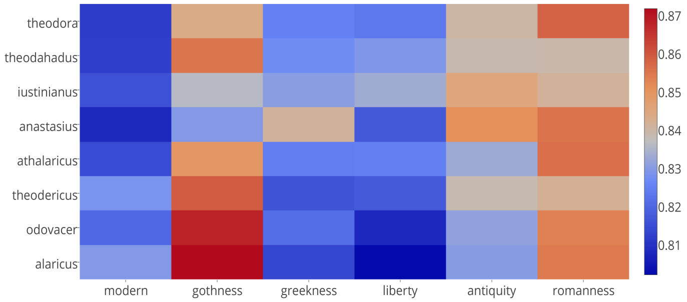


Figure 2: Reproduction de l'expérience menée par Bjerva et. al., avec notre modèle modifié.

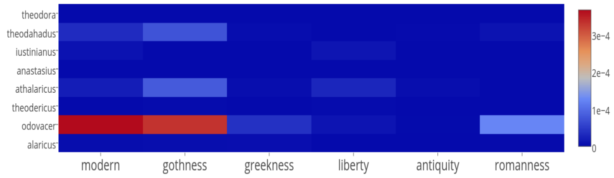


Figure 3: Comparaison des similarités Personne/Concept entre le modèle de Bjerva et. al. et notre modèle modifié. Chaque cellule représente la valeur absolue de la différence de similarité entre les deux modèles. Les cellules rouges présentent le plus de différences.

Nous avons expérimenté notre méthode en reproduisant l'expérience menée par Bjerva et. al. (Bjerva, 2015). Ces derniers se sont intéressés aux relations qu'entretiennent différentes personnalités du VI^{ème} siècle avec de grands concepts (Modernité, Liberté, Gothique, ...). Ils ont utilisé Word2Vec pour entraîner un modèle sur environ 11 000 textes latins, pour ensuite comparer les distances qui séparent les personnes des concepts dans l'espace vectoriel du modèle (voir figure 1). Nous avons utilisé notre méthode pour calculer, pour chaque personne d'intérêt, un nouveau vecteur représentatif prenant en compte les différentes erreurs OCR identifiées. Les distances entre personnes et concepts au sein de notre modèle modifié sont présentées dans la figure 2. Pour plus de clarté, les deux modèles sont comparés dans la figure 3. On peut par exemple observer qu'Odovacer, la personne pour qui les différences sont les plus grandes, est assez peu citée dans le corpus. Notre méthode a cependant identifié de nombreuses erreurs OCR qui ont révélé des informations inconnues au seul vecteur de l'entité originale.

La méthode présentée ici permet d'identifier de potentielles erreurs OCR sur les entités nommées au sein d'un corpus. La prise en compte de ces erreurs peut avoir un impact non négligeable sur le modèle et donc sur les analyses qui en découlent. Cela semble en particulier vrai pour les entités nommées peu présentes dans un corpus.

Bibliographie

Bjerva, J. and Praet, R. (2015). "Word Embeddings Pointing the Way for Late Antiquity." LaTeCH 2015: 53.

Gefen, A. (2015). Les enjeux épistémologiques des humanités numériques. Socio. La nouvelle revue des sciences sociales, 4: 61-74.

Hamilton, W. (2016). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1.

Lemercier, C. (2005). Analyse de réseaux et histoire. Revue D'histoire Moderne et Contemporaine(2): 88–112.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems. pp. 3111–3119

Piotrowski, M. (2012). Natural Language Processing for Historical Texts. (Synthesis Lectures on Human Language Technologies 17). San Rafael, CA: Morgan & Claypool.

Body Parts in Norwegian Books

Lars Gunnarsønn Johnsen
yoonsen@gmail.com
National Library of Norway, Norway

Introduction

The goal of the present study is to look at body parts within the class of published Norwegian books. Our main question is to look at the difference between referencing a body part directly from making a reference via a possessive pronoun. The body is important in our culture and is studied both within and outside the digital realm, e.g. the papers in Christopher E Forth; Ivan Crozier (2005) studies the body in culture, while Mahlberg (2013) uses corpus methods in a literary study.

We report on a pilot study that considers body parts across the whole book collection without breaking the collection up into different genres or time periods. Although the pilot study is limited to an across-the-books analysis, it is part of effort to study the effect of different genres, as well as newspapers and journals.

Method and data

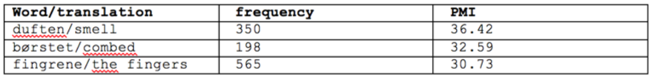
Words for body parts are taken from Norwegian books in the period 1810 up to 2000, using the digitized books made available through the Norwegian National Library, approximately 460 000 books.

The contexts we consider for nouns describing body parts fall into three types as described in Lødrup (2011) and Delsing (1998). Norwegian may express possessives parallel to the English pattern, like "hans arm (his arm)", alongside the definite version like "armene hans (arms.PL.DEF his)". Modern Norwegian seems to prefer the definite plus possessive, in particular for body parts, which therefore will be the construction we focus mostly on here.

A slight complicating factor in Norwegian possessive construction is that sometimes the possessor may not be expressed e.g. Lødrup (op.cit.), in contrast to English, where the possessor cannot easily be replaced with "the" in "John had a pain in his arm" while in Norwegian this is the norm if the possessor is the subject "John hadde vondt i armen/John had pain in the arm".

Each possessor phrase, pronoun plus body part gets a collocation graph as described in Brezina et.al (2015) where each edge in the graph is weighted with PMI (Pointwise Mutual Information, e.g. Lewandowska-Tomaszczyk (2007), Romesburg (2004)). Collocation graphs can be seen as a cluster of words for the phrase generating it, ordered by PMI.

As an example, the first three words in the cluster (or collocation graph) for "håret hennes (her hair)", computed from approximately 12 000 concordance samples, go like this:



Word/translation	frequency	PMI
duften/smell	350	36.42
børstet/combed	198	32.59
fingrene/the fingers	565	30.73

There are 29 Norwegian body words in all going into this study, some duplicated in singular and plural, resulting in 21 unique body parts from head to toe.

Each cluster is cut down to its top 200 words which are then compared using two standard similarity measures, the cosine and the Jaccard-similarity, where the former accentuates similarity of the clusters as weighted distributions, the latter highlights the set equality of the clusters.

Results

Our research question is how references to body parts differ when referenced using a pronoun, or with no pronoun specified. Note that even though no pronoun is specified, it is not required that the reference is done without a possessor, so there will be a certain overlap in the samples.

Our main result is that female and male possessive constructions generate clusters that are closer than standalone body words. Looking at the clusters themselves we see that some is due to the expressiveness of the body, for example "øynene/the eyes", which ranks on top between female and male parts, has words like "glimt/sparkle" and "lyste/lightened" which is used to express emotion emanating from the beholder. These words are absent for the cluster for the word "øyne"eyes". Also, words like "hendene/the hands" in the possessive construction yield words of action like "grep/gripped", "slapp/released", while outside the possessor construction generic actions like "klappe/clap" is found.

The next step is to study these constructions with respect to a distinction between classes of works, using the metadata of national bibliography, and also the difference across media types such as newspapers, books and journals.

Bibliography

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173.

Delsing , L. O. (1998) "Possession in Germanic", in *Possessors, Predicates and Movement in the Determiner Phrase*, ed. Artemis Alexiadou & Chris Wilder, John Benjamins, Amsterdam

Forth, C E, and Crozier , I., eds. (2005) *Body parts : critical explorations in corporeality* Lanham : Lexington Books.

Lewandowska-Tomaszczyk , B (2007) *Corpus linguistics, computer tools, and applications : state of the art* P.Lang Frankfurt am Main, New York

Lødrup, H (2011), "Norwegian Possessive Pronouns: Phrases, Words or Suffixes?" in *Proceedings of the LFG11 Conference*, Butt, Miriam and King Tracy eds., CSLI Publications, http://csli-publications.stanford.edu/

Mahlberg, M (2013). *Corpus Stylistics and Dicken's Fiction*. Routledge.

Romesburg , H. C. (2004) *Cluster analysis for researchers*, Lulu Pr.

Constructing Linked Knowledge around Southeast Asian Studies

Akihiro Kameda
kameda@cseas.kyoto-u.ac.jp
Center for Integrated Area Studies
Kyoto University, Japan

Shoichiro Hara
shara@cseas.kyoto-u.ac.jp
Center for Integrated Area Studies
Kyoto University, Japan

To understand an academic humanities paper, it is crucial to understand each element in the paper (person, place, mention of another paper, etc.). To understand each element in the paper, we have to refer to knowledge outside of the paper; in other words, linked knowledge is important. Linked Open Data (LOD) (Heather and Bizer, 2011), which is emerging from research in the Semantic Web, provides the way to represent and share such linked knowledge.

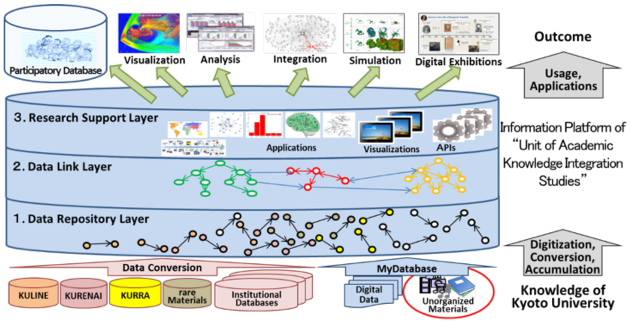
We prepared texts of "Japanese Journal of Southeast Asian Studies" as a core dataset, in order to attempt to link words or documents in the core dataset to external resources such as DBpedia (also available in Japanese) or the National Diet Library in Japan.

Though LOD has been growing rapidly (Schmachtenberg, 2014), it is difficult to cover specific knowledge in each academic paper. Therefore, the publication of LOD is also an important effort to represent knowledge networks in academic humanities papers.

The Center for Integrated Area Studies, Kyoto University (CIAS) has developed two information tools, named MyDatabase (MyDB) and Resource Sharing System (RSS), to solve these difficulties. The main component of MyDB is a database builder, allowing humanities researchers to construct and revise databases without expert knowledge. MyDB stores metadata and accepts any vocabulary of metadata, including nonstandard vocabularies. This enables humanities researchers to use their own metadata vocabulary according to their own purposes. On the other hand, those metadata varieties make the integration processes difficult. RSS was developed to integrate heterogeneous databases on the Internet, and to provide users with a uniform interface to retrieve databases seamlessly in one operation. Thus, MyDB and RSS have helped accelerate open data in the humanities. However, there are still two problems to solve, especially in the case of RSS: limited coverage of databases and initial costs of integration. First, for example, Kyoto University released KULINE (OPAC), KURENAI (repository), KURRA (archive), Open Course Ware and various databases developed by each research institute in the university, but RSS does not integrate these databases. Second, it is time consuming to integrate new databases into RSS and impossible to trace links automatically. As such, for now, RSS is not the appropriate tool to discover hints and/or create new knowledge.

To overcome these drawbacks, a new project has been launched to develop an innovative information platform for open humanities data. This platform comprises three sublayers. The first layer is "Open Data Layer" which accumulates heterogeneous metadata. This layer uses RDF to describe data of different structures. The second layer is "Data Link Layer." This layer uses ontology techniques such as RDFS and OWL to link ambiguous (uncontrolled) vocabularies and emerge "humanities big data." The third layer is "Application Layer." As big data in the humanities is too huge and complicated to retrieve, categorize, and analyze by hand, this layer provides utilities to process big data. This platform will prepare for APIs to help mashup applications. We expect the platform to reconstruct a knowledgebase from heterogeneous databases, which is used to construct meaningful chunks from scattered data.

Thus, humanities Linked Open Data has been developed, and the "Japanese Journal of Southeast Asian Studies" dataset can be linked to that LOD. This linked knowledge can then help readers from other domains.



Bibliography

Heath, T. and Bizer, Ch. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.

Schmachtenberg, M. et al. (2014). *Linking Open Data cloud diagram* http://lod-cloud.net/

DH Box

Jojo Karlin
jkarlin@gradcenter.cuny.edu
The Graduate Center, CUNY, United States of America

Patrick Smyth
patricksmyth01@gmail.com
The Graduate Center, CUNY, United States of America

Stephen Zweibel
szweibel@gc.cuny.edu
The Graduate Center, CUNY, United States of America

Matthew K. Gold
mgold@gc.cuny.edu
The Graduate Center, CUNY, United States of America

This poster session will introduce DH Box, a browser-based platform that provides access to a variety of difficult-to-install digital humanities tools. Currently under development at The Graduate Center, CUNY, and funded by generous assistance from an NEH Startup Grant, DH Box is designed with particular attention to digital humanities pedagogy. Because teaching DH requires infrastructure of various kinds, including computing resources, IT support, and teachers with specialized training, student access to the digital humanities is unevenly distributed and highly dependent on local institutional conditions. Faculty with DH skills are often called upon to teach classes and workshops in adverse circumstances, contending with limited access to labs or other computing resources, restrictive IT policies that prevent new software from being installed, and platform fragmentation on student-owned machines. As a cloud-based platform, DH Box was conceived to address some of these concerns by providing a set of digital humanities tools through a unified computing environment accessed through the browser.

DH Box is not a service, nor is it a standalone application that runs on local computers. Rather, it is software that can be set up by teachers or institutions on cloud-based infrastructure such as Amazon Web Services or DigitalOcean. DH Box is an open-source project developed by and for teachers and researchers in the humanities, and as such is intended to provide an alternative to proprietary services that prevent access to user data. DH Box also makes an intervention in terms of design, UX, and usability. The platform uses a virtualization technology called Docker to quickly create and remove digital working environments, providing students with access to either shared work spaces or individualized virtual machines. Each of these environments contains a set of tools and utilities frequently used in the digital humanities, including IPython Notebooks, the Natural Language Toolkit, MALLET, Omeka, and WordPress. DHBox's tab-based browsing interface makes it possible to easily switch between these utilities, the command line, and a text editor. In addition to providing DH tools in a unified computing environment, the platform offers resources in partnership with institutions like the British Library, making digitized texts available for student exploration. By integrating Git Lit, a tool for downloading text corpora developed in partnership with Columbia University, DH Box provides access not only to digital humanities tools, but also to materials for analysis and experimentation.

Even as it aims to circumvent institutional barriers to DH scholarship, DH Box addresses broader questions of access to DH tools. The necessity for specialized knowledge--use of the command line, importing packages and modules, configuring a working environment--may deter students from pursuing research questions using these novel methods. Making these tools more accessible, on the other hand, should help encourage a new generation of DH scholars. By allowing teachers and students to bypass the difficult process of installation and configuration, the DH Box team hopes to give them room to focus on exploration and experimentation. During the DH Box poster presentation, team members will be available to discuss decisions of design as well as future use cases.

Resources

For more information about DH Box, please refer to: Dhbox.org | @DH_Box | https://github.com/DH-Box

Unpacking Collaboration in Digital History Projects

Max Kemman
max.kemman@uni.lu
University of Luxembourg, Luxembourg

Digital history is concerned with the incorporation of digital methods in historical research practices. Thus, digital history aims to use methods, concepts, or tools from other disciplines to the benefit of historical research, making it a form of **methodological interdisciplinarity** (Klein, 2014). This requires expertise of different facets, such as technology, history, and data management, and as a result many digital history activities are a collaboration of professionals and scholars from different backgrounds. Such collaborations would fit Svensson's characterisation of digital humanities as **a fractioned trading zone** (2011; 2012). Simply stated, this means first that digital humanities functions as heterogeneous collaborations, i.e. with participants from different backgrounds, and second that the participants act voluntarily. In this paper, we will investigate these two aspects in the context of digital history to understand how digital history projects function as heterogeneous collaborations, and what the participants' incentives are for entering such collaborations. We will discuss this by presenting findings from interviews with practitioners in digital history projects, and reflections on projects in which the author himself has participated.

The concept of **trading zones** was coined by Galison who described it as "an arena in which radically different activities could be locally, but not globally, coordinated" (1996, p. 119). That is, although the disciplines of e.g. computer science and history cannot coordinate activities on a global discipline-wide level, and do not contribute towards one another as disciplines, in local collaborations it is possible to communicate and coordinate a shared goal of research within a so-called trading zone. This concept was further developed by Collins et al. (2007) who suggested four types of trading zones using two dimensions. The first dimension is **cultural maintenance** from homogeneous to heterogeneous, i.e., how the two groups define themselves and to what extent they aim to maintain their identity. On this scale, more homogeneous means the two groups become more alike to form a single group, while more heterogeneous means they remain two distinct groups. The second is **coercion** from collaborative to coercive, i.e., what the power relations in the trading zone are. On this scale, more collaborative means the two groups are both acting out of free will, while more coercive means one group is imposing practices upon the other. When a trading zone is heterogeneous and collaborative, we speak of a fractioned trading zone as Svensson does.

One instantiation of this is through **boundary objects**, a concept developed by Star and Griesemer to describe objects used in heterogeneous collaborations where different parties may have different understandings of the object, while the object keeps a common core identity to all parties (Star and Griesemer, 1989; Star, 2010). This concept could be used to refer to the tool under development or the data on which the tool and historian will work. However, in this paper we will approach the project itself as boundary object; the project binds the participants together, but we will ask what each participant expects out of the project, and how participants individually approach the project.

This leads us to the second part of our investigation, the **incentives** for collaboration. When writing about interdisciplinary collaboration in digital history, this is almost always done to underscore the positive or even necessary effects (e.g. Eijnatten et al., 2013; Hitchcock, 2014; Sternfeld, 2011). However, such collaboration is not trivial and requires dedication and investments from all involved, e.g. as shown by Siemens (Siemens et al., 2009; Siemens and INKE Research Group, 2012). In previous research, it has been shown that the incentives for joining a project had a strong influence on the success of collaborations between computer scientists and earth scientists (Weedman, 1998). To understand these incentives, we follow this work and look at reasons for joining the project, individual goals for the project, and expected effects of the participation after the project has ended. From these aspects, we will analyse situations of conflicting interests and expectations. For example, in an interview one historian noted about their project:

> ”[W]e’re supposed to be advising the team developing the tool. And trying to then carry out research on a specific case study. And so originally it was like wow we’re going to be able to use the tool, but very quickly it became clear ok actually probably we’re not going to be able to use the tool.”

In this paper, we will thus unpack the fractioned trading zones of digital history projects, to gain an understanding of how heterogeneous, interdisciplinary collaborations work, and why participants join these collaborations.

Bibliography

Collins, H., Evans, R. and Gorman, M. (2007). Trading zones and interactional expertise. *Studies in History and Philosophy of Science* Part A, 38(4): 657–66 doi:10.1016/j.shpsa.2007.09.003.

Eijnatten, J. van, Pieters, T. and Verheul, J. (2013). Big Data for Global History: The Transformative Promise of Digital Humanities. BMGN - *Low Countries Historical Review*, 128(4): 55–77.

Galison, P. (1996). Computer simulations and the trading zone. The Disunity of Science: Boundaries, Contexts, And Power. Stanford University Press, pp. 118–57.

Hitchcock, T. (2014). Big Data, Small Data and Meaning Historyonics http://historyonics.blogspot.co.uk/2014/11/big-data-small-data-and-meaning_9.html.

Klein, J. T. (2014). Interdisciplining Digital Humanities: Boundary Work in an Emerging Field. online. University of Michigan Press doi:10.3998/dh.12869322.0001.001.

Siemens, L., Duff, W., Cunningham, R. and Warwick, C. (2009). “It challenges members to think of their work through another kind of specialist”s eyes’: Exploration of the benefits and challenges of diversity in digital project teams. Proceedings of the American Society for Information Science and Technology, 46(1): 1–14 doi:10.1002/meet.2009.1450460223.

Siemens, L. and INKE Research Group (2012). From Writing the Grant to Working the Grant : An Exploration of Processes and Procedures in Transition. Scholarly and Research Communication, 3(1).

Star, S. L. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. Science, Technology & Human Values, 35(5): 601–17 doi:10.1177/0162243910377624.

Star, S. L. and Griesemer, J. R. (1989). Institutional Ecology, `Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. Social Studies of Science, 19(3): 387–420 doi:10.1177/030631289019003001.

Sternfeld, J. (2011). Archival theory and digital historiography: Selection, search, and metadata as archival processes for assessing historical contextualization. American Archivist, 74(2): 544–75.

Svensson, P. (2011). The digital humanities as a humanities project. *a* 11(1–2): 42–60 doi:10.1177/1474022211427367.

Svensson, P. (2012). Beyond the Big Tent. In Gold, M. K. (ed), *Debates in the Digital Humanities.* online. University of Minnesota Press.

Weedman, J. (1998). The Structure of Incentive: Design and Client Roles in Application-Oriented Research. Science, *Technology & Human Values,* 23(3): 315–45 doi:10.1177/016224399802300303.

MemoryGraph: Digital Critique of Old Photographs Using a Mobile App that Enhances the Interpretation of Landscape

Asanobou Kitamoto
kitamoto@nii.ac.jp
National Institute of Informatics, Japan

Introduction

MemoryGraph is a medium to record memories as the augmentation of a photograph, which is a medium to record photons. MemoryGraph is a new photographic technique that creates a layer of “memories” (or more precisely, time-series photographs) which are taken by the same composition at different times. Photographs of the same composition can be taken by traditional cameras, but this requires substantial effort, because matching a printed photograph with the current landscape requires mental rotation (Shepard et al., 1971), a psychologically demanding task. MemoryGraph simplifies this task through direct overlay of a photograph and the current landscape on a camera viewfinder with adjustable transparency. We demonstrate that this simple idea opens up new possibilities for the critical interpretation of photographs in the context of digital critique.

Related Methods

MemoryGraph focuses on the spatio-temporal relationship between photographs and the real world. This field of interest has given rise to many methods similar to MemoryGraph. For this reason, we begin with a comparison between MemoryGraph and these related methods in order to characterize MemoryGraph’s unique properties.

First, it is necessary to compare MemoryGraph with time-lapse animation, which is also a time-series image of the same composition. The fundamental difference between time-lapse animation and MemoryGraph is in time scale and device dependence. Time lapse animation deals with high frequency observations of seconds to minutes using the same location-fized camera, while MemoryGraph deals with low frequency observations over a period of days to years. These observations may potentially use different cameras that are not fixed at the location. In short, MemoryGraph is a tool to realize fixed point observation at any place for any time interval.

Secondly, we contrast MemoryGraph with augmented reality (AR), which focuses on the alignment of the real space and the virtual space so that a photograph can be seen as an overlay on the real space through a camera viewfinder. On the contrary, MemoryGraph focuses on the overlay of a photograph on a camera viewfinder as a graphic reference to illustrate the composition of a further action (namely, taking a photograph). This key contrast suggests fundamentally different roles between augmented reality and MemoryGraph. Augmented reality is a tool for exhibition, in the sense that alignment is controlled by a tool, while the user is allowed to be a passive visitor who experiences an environment prepared by someone else. Quite the opposite, MemoryGraph is a tool for participation: the "visitor" controls the alignment, and the user should be an active explorer who searches for the best match between photograph and real-life landscape. In short, MemoryGraph is a tool for actions, such as participatory annotation.

Our final comparison discusses photo-sharing services dedicated to old photographs. For example, Historypin (Shift, 2010) is designed to share photographs of the past using a map interface, while an advanced system might use a street-view interface to place photographs in 3D. Both tools aim to link photographs to the real world, however, their methods of achieving this goal do not rely on ventures into the "real world" to "place" the photographs. Photo sharing services depend on "arm-chair" annotators, but MemoryGraph depends on "field" annotators who visit a real place to take another photograph. In short, MemoryGraph is a tool to motivate people to move in the real world.

Proposed Method

MemoryGraph is designed as a mobile app for two reasons. First, the idea of "graphic reference" overlay on a camera viewfinder cannot be implemented on traditional cameras that do not expose API (application programming interface). By re-purposing the viewfinder of a smart phone, MemoryGraph can extend the grid-based reference of a traditional camera viewfinder to a graphic reference such as an old photograph. Secondly, in order employ MemoryGraph as a field work tool, the use of a mobile phone is the best choice for mobility and also for real-time information sharing with the server.

The task of the user is to find the best match between a graphic reference and the real world using an opaque viewfinder with adjustable transparency. Direct comparison between two scenes not only reduces the burden of mental rotation for the user, but also makes the search enjoyable: the movement of the user gives real-time visual feedback that suggests how "good" the move is. This gamification effect motivates users to search for better matching, and promotes photographic crowd-sourcing for participatory annotation.

The outcome of this task is two types of data. One is a photograph that records the current landscape, and another is the metadata of the photograph such as latitude, longitude, and direction observed by sensors in a mobile phone. Metadata may be enhanced later to add a title, and description (among other things), and users can upload metadata and photographs to the server for sharing. The uploaded data can then be used for scholarly research, because the GPS coordinates and the temporally different views of the landscape are valuable resources for understanding the landscape changes.

Results

The MemoryGraph (CODH, 2016) app is available for free on Google Play, but an iOS version has not yet been released. The predecessor app, MemoryHunt (Kitamoto, 2015), has been used for both the study of cultural landscapes and the monitoring of disaster recovery (DSR, 2014).

To study cultural landscapes, we held several workshops with both scholars and laypersons who tried the app in the field. An example is shown in Figure 1, where the reference image was a photograph of Imperial Palace Moat in Tokyo. The photograph was easy to interpret, but the actual place was difficult to find. Figure 2 shows how participants walked along the moat to find the best match. At each place, participants took the photograph that they believed to be the best match, but the best solution of Figure 1 was found after many trials of all the participants.

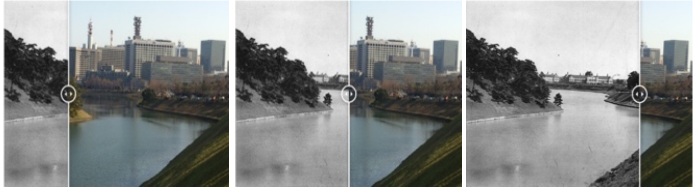


Figure 1. Comparison of two photographs of the same composition. "A view from Miyakezaka to the ministry of justice" taken around 1911 and the current landscape. Courtesy of National Diet Library.

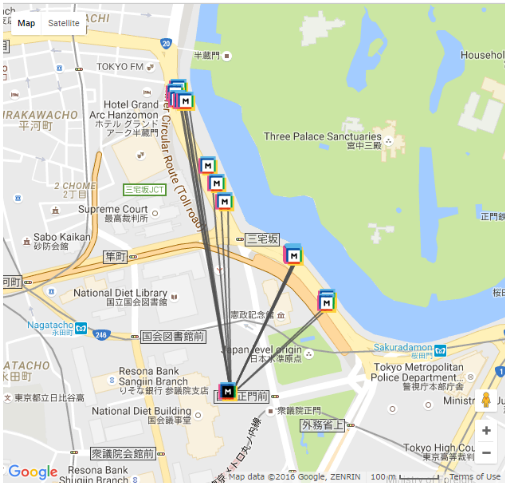


Figure 2. Participants walked along the moat to find the best match, and their trials are marked by icons. The final solution of Figure 1 was obtained at the north end of the moat.

For the second purpose, we held workshops at Kobe in Japan, which was severely damaged by the earthquake in 1995, and Aceh in Indonesia, which was severely damaged by the tsunami in 2004, to understand how the city recovered from the disaster. The app was enjoyed by local people in two countries, and some children were involved more actively than adults due to the gamification effect.

MemoryGraph can be generalized in several ways. First, it can be generalized from landscape to object: for example, time-series photographs of the same person at different places. Second, it can be generalized to cross-media reference: for example, taking the same composition at a "sacred place" of pop-culture work inspired by the real landscape.

Discussion on Digital Critique

Digital critique, or what has been referred to in other words as digital criticism (Kitamoto, 2016) or data criticism (Kitamoto, 2014), was proposed by the author as a framework for digital scholarship in the humanities. It has been applied to the criticism of non-textural sources such as maps and photographs with the intent of using them as evidence for historical studies. Although the digital humanities are often considered quantitative or metric humanities, digital critique focuses on the digital creation and management of humanities knowledge. Several types of digital tools have been designed for the critical interpretation of textual and non-textural sources, including success stories such as the identification of Silk Road ruins, or the characterization of mistakes in the renovation of old Beijing maps (Kitamoto et al., 2014) (Kitamoto et al., 2016). We claim that MemoryGraph is another digital critique tool that asks the following research questions in order to use a photographic source as historical evidence.

The first research question asks what the variance and invariance is in the target landscape over time. Invariance is often found as artificial features such as roads and monuments, or natural features such as a mountain ridge, but in other cases, the identification of invariance requires training of the user in terms of interpretive performance on the historical landscape.

Secondly, we ask how historical evidences can be best integrated across different sources, such as old maps, old photographs and historical databases. Invariance found through the app is a hint to links different sources over time, and this contributes to expanding our understanding by integrating knowledge from different sources.

Our final research question asks why the photograph was taken in this setting. This is because the best match is theoretically achieved by the same physical posture between the original photographer and the user (Kitamoto, 2015). This means that as a user tries to find the best match using the app, a user can re-experience the original photographer who took the photograph in the same posture. The physical overlay of the user and the photographer may lead to research questions investigating the emotion or the way of thinking of the original photographer at the time. This has potential to give birth to new interpretation of photographs from the physical or emotional perspective.

Acknowledgments

The project was partially supported by JSPS Kakenhi Center Grant Number 16H02920 and by Center for Integrated Area Studies (CIAS), Kyoto University. The core contributor to the app is Mr. Tomohiro Ikezaki.

Bibliography

Center for Open Data in the Humanities (CODH) (2016). "MemoryGraph." https://mp.ex.nii.ac.jp/mg/, (accessed on April 6, 2017).

Digital Silk Road Project (DSR) (2014). "MemoryHunt." http://dsr.nii.ac.jp/memory-hunting/, (accessed on April 6, 2017).

Kitamoto, A., and Nishimura, Y. (2016). "Digital criticism platform for evidence-based digital humanities with applications to historical studies of Silk Road." Digital Humanities 2016:Conference Abstracts.

Kitamoto, A., and Nishimura, Y. (2014). "Data criticism: general framework for the quantitative interpretation of non-textual sources." Digital Humanities 2014:Conference Abstracts.

Kitamoto, A. (2015). "MemoryHunt: a mobile app with an active viewfinder for crowdsourced annotation through the re-experience of the photographer", Fifth Annual Conference of the Japanese Association for Digital Humanities: Conference Abstracts.

Shepard, R.N., Metzler, J. (1971). "Mental rotation of three-dimensional objects." Science, 171(3972): 701-703.

Shift (2010). "Historypin." http://www.historypin.org/, (accessed on April 6, 2017).

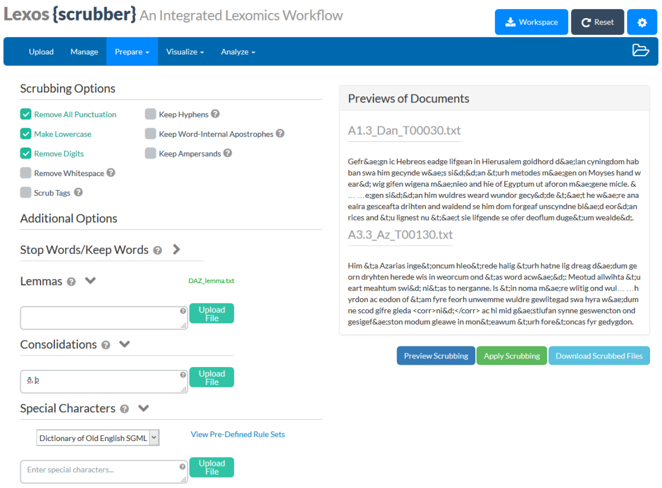
Lexos: An Integrated Lexomics Workflow

Scott Kleinman
scott.kleinman@csun.edu
California State University: Northridge, United States

Mark LeBlanc
mleblanc@wheatoncollege.edu
Wheaton College, United States

Lexos is a browser-based suite of tools that helps lower barriers of entry to computational text analysis for humanities scholars and students. Situated within a clean and simple interface, *Lexos* consolidates the common pre-processing operations needed for subsequent analysis, either with *Lexos* or with external tools. It is especially useful for scholars who wish to engage in research involving computational text analysis and/or wish to teach their students how to do so but lack the time for a manual preparation of texts, the skill sets needed to prepare their texts analysis, or the intellectual contexts for situating computational methods within their work. *Lexos* is also targeted at researchers studying early texts and texts in non-Western languages, which may involve specialized processing rules. It is thus designed to facilitate advanced research in these fields even for users more familiar with computational techniques. *Lexos* is developed by the Lexomics research group led by Michael Drout (Wheaton College), Mark LeBlanc (Wheaton College), and Scott Kleinman (California State University, Northridge). It is built on Python 2.7-Flask microframework, with jQuery-Bootstrap UI, and visualizations in d3.js. The Lexomics research group provides access to an public installation of *Lexos* which does not retain data after a session has expired. Users may also install *Lexos* locally by cloning the GitHub repository.

Lexos guides users through a workflow of steps that reflects effective practices when working with digitized texts. The workflow includes: (i) uploading Unicode-encoded texts in plain text, HTML, or XML formats; (ii) "scrubbing" functions for consolidating pre-processing decisions such as the handling of punctuation, white-space, and stop words, the use of lemmatization rules, and the handling of embedded markup tags and special character entities; (iii) "cutting" texts into segments based on the number of characters, tokens, or lines, or by embedded milestones such as chapter breaks; (iv) tokenization into a Document Term Matrix of raw or proportional counts using character or word n-grams; (v) visualizations such as comparative word clouds per segment (including the ability to visualize topic models generated by MALLET); Rolling Window Analysis that plots the frequency of string, phrase, or regular expression patterns or pattern-pair ratios over the course of a document or collection; and (vi) analysis tools including statistical summaries, hierarchical and k-means clustering, cosine similarity rankings, and Z-tests to identify the relative prominence of terms in documents, document classes, and the collection as whole. At each stage in the workflow the user may download data, visualizations, or the results of the analytical tools, along with metadata about their pre-processing decisions or the parameters selected for their experiments. *Lexos* thus enables the export of data for use with other tools and facilitates experimental reproducibility.



Scrubbing Options

☑ Remove All Punctuation
☑ Make Lowercase
☑ Remove Digits
□ Remove Whitespace
□ Scrub Tags
□ Keep Hyphens
□ Keep Word-Internal Apostrophes
□ Keep Ampersands

Additional Options

Stop Words/Keep Words

Lemmas: DAZ_lemma.txt

Consolidations: ð, þ

Special Characters: Dictionary of Old English SGML

Enter special characters...

Previews of Documents

A1.3_Dan_T00030.txt

Gefr&ae;gn ic Hebreos eadge lifgean in Hierusalem goldhord d&ae;lan cyningdom hab ban swa him gecynde w&ae;s si&d;&d;an &t;urh metodes m&ae;gen on Moyses hand w ear&d; wig gifen wigena m&ae;nieo and hie of Egyptum ut aforon m&ae;gene micle. & … …e;gen si&d;&d;an him wuldres weard wundor gecy&d;de &t;&ae;t he w&ae;re ana ealra gesceafta drihten and waldend se him dom forgeaf unscyndne bl&ae;d eor&d;an rices and &t;u lignest nu &t;&ae;t sie lifgende se ofer deoflum duge&t;um wealde&d;.

A3.3_Az_T00130.txt

Him &t;a Azarias inge&t;oncum hleo&t;rede halig &t;urh hatne lig dreag d&ae;dum ge orn dryhten herede wis in weorcum ond &t;as word acw&ae;&d;: Meotud allwihta &t;u eart meahtum swi&d; ni&t;as to nerganne. Is &t;in noma m&ae;re wlitig ond wul… …h yrdon ac eodon of &t;am fyre feorh unwemme wuldre gewlitegad swa hyra w&ae;dum ne scod gifre gleda <corr>ni&d;</corr> ac hi mid g&ae;stlufan synne geswencton ond gesigef&ae;ston modum gleawe in mon&t;eawum &t;urh fore&t;oncas fyr gedygdon.

Figure 1: The *Lexos* Scrubber Tool

Lexos addresses three significant challenges for our intended users. The first challenge involves the **adoption** of computational text analysis methods. Many approaches require proficiency with command line scripting or the use of complex user interfaces that require time to master. *Lexos* addresses this problem through a simple, browser-based interface that manages workflow through the three major steps of text analysis: pre-processing, generation of statistical data, and visualization. In this, *Lexos* resembles *Voyant Tools* (Sinclair and Rockwell, 2016), although *Lexos* places more emphasis on and providing more tools for pre-processing and segmenting texts. *Lexos* also shares with tools like *Stylometry with R* (Eder, et al., 2013; Eder, 2013) and emphasis on cluster analysis, providing both hierarchical and K-Means clustering with silhouette scores as limited form of statistical validation. While *Lexos* is not a topic modeling tool, it provides a useful "topic cloud" feature for MALLET data that will be useful for beginners since there are few accessible ways to visualize MALLET output that work well out of the box.

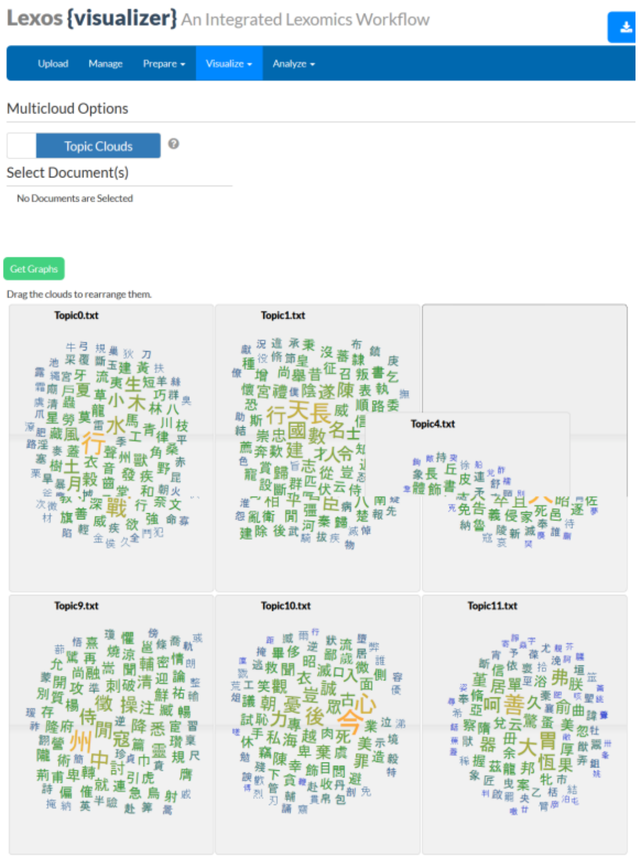


Figure 2: The *Lexos* Multicloud tool showing Chinese "topic clouds"

The second challenge is the **opacity** of the procedures required to move between computational and traditional forms of text analysis. In order to reduce the "black boxiness" of algorithmic methods, *Lexos* contains an embedded component called "In the Margins" which provides non-technical explanations of the statistical methods used and effective practices for handling situations typical of humanities data. "In the Margins" is a Scalar "book" which can be read separately; however, its individual pages are embedded in *Lexos* using Scalar's API, making them easily accessible for users of the tool. *Lexos* shares with tools like *Voyant* an engagement with the hermeneutics of text analysis and attempts to embed "In the Margins" discussion of these issues in the user interface close to the user's workflow. We hope "In the Margins" will host advice and commentary from contributors with the Digital Humanities community.

A third challenge is the **tension** between quantitative and computational approaches and the traditions of theoretical and cultural criticism that dominate the humanities in the academy. As Alan Liu (2013) has recently argued, the challenge is to give a better theoretical grounding to the hybrid quantitative-qualitative method of the Digital Humanities by exploring the ways in which we negotiate the difficulties imposed by "the aporia between tabula rasa quantitative interpretation and humanly meaningful qualitative interpretation" (414). The design of *Lexos* and the discussions in "In the Margins" are intended to open a space for discussion of issues related to the opacity of algorithmic approaches and the limitations and epistemological challenges of computational stylistic analysis and visual representation of humanities data.

This poster presentation provides demonstrations of *Lexos* using some literature from Old, Middle, and Modern English, as well Chinese, which are in our current test suite. We also discuss use cases and best practices, how to install *Lexos* locally, and how scholars may contribute to the still growing content of "In the Margins".

Bibliography

Drout, M., Kleinman, S., and LeBlanc, M. 2016-. "In the Margins." http://scalar.usc.edu/works/lexos./

Eder, M. (2013). "Mind Your Corpus: Systematic Errors in Authorship Attribution." *Literary and Linguistic Computing* 28 (4): 603–14.

Eder, M., Kestemont, M., and Rybiki, J. 2013. "Stylometry with R: A Suite of Tools (Abstract of Poster Session)". Presented at Digital Humanities 2013, Lincoln, Nebraska. http://dh2013.unl.edu/abstracts/ab-136.html, https://sites.google.com/site/computationalstylistics/

Kleinman, S., LeBlanc, M.D., Drout, M. and Zhang, C. 2016. *Lexos* v3.0. https://github.com/WheatonCS/Lexos/.

Liu, A. (2013). "The Meaning of the Digital Humanities." *PMLA* 128 (2): 409-23.

McCallum, A.K. (2002). *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Sinclair, S., and Rockwell, G. (2016). *Voyant Tools*. Web. http://voyant-tools.org/.

Collaborative Approaches to Open up Russian Manuscript Lexicons

Kira Kovalenko
kira.kovalenko@gmail.com
Austrian Academy of Sciences, Austria

Eveline Wandl-Vogt
eveline.wandl-vogt@oeaw.ac.at
Austrian Academy of Sciences, Austria

In this paper, the authors introduce a new research collaboration between the Institute for Linguistic Studies (St. Petersburg, Russia) and the Austrian Academy of Sciences, Austrian Centre for Digital Humanities (Vienna, Austria), and exemplify multicultural, crossdisciplinary collaboration on the Ph.D project of Russian manuscript lexicons.

The project is aimed at the online representation of the Russian manuscript lexicons, which reflect a very important stage in Russian lexicography and served as a basis for the printed dictionaries appeared later. This lexicographical genre (called azbukovnik in Russian) appeared in the middle of the 16th century on the base of glossaries, and differ from them by the alphabetical order of the word entries. About 150 manuscripts, mostly from the 17th century, containing the lexicons came into our hands. The tradition of copying manuscript lexicons preserved in the Old Believers communities, and the latest lexicon dates from the beginning of the 20th century. Lexicons were classified by L.S.Kovtun in the 1989, and some additions were made later by K.I.Kovalenko (for details see Kovalenko, 2016).

Obligatory parts of the word entries were head word (or collocation) and explanation. Word entries could also contain inscription of the language source for loan words, lexicographical or literary sources, citations, and information about topic groups (animals, birds, plants, etc.). Sometimes, observations on various linguistic topics were included in the word entries as well. Despite the great importance of these lexicons as Russian language history resources, only small part of them is available for researchers (2 published in Kovtun, 1975 and 1989, and 16 can be found online as a facsimile.

The research idea is formulated collaboratively by the Russian and Austrian scholars, experienced in (traditional) lexicography as well as in digital transformation and research infrastructure development. An architecture to embed the Russian manuscript lexicons into a larger lexical resources infrastructure is developed based on these visions and needs at the Austrian Academy of Sciences. Cornerstones of the architecture are opening up lexical resources and develop as much as possible (due to different types of licensing) freely available resources to the broader public which meet the open science definition.

The established architecture consists of three layers:

1. **The human interface layer:** supports data access e.g. list of headwords, geographical map, timeline and other visual approaches etc., which is further developed with every new project. Furthermore, via authentification it offers a private workspace.
2. **The persistent layer:** offers a data repository, e.g. triple store, open access repository to sustainable store and interlink data.
3. **The enrichment layer**: offers opportunities and tools to interlink data and collaboratively enrich data, e.g. with Wikidata or Europeana.

Ideally, all Russian manuscript lexicons in text format and as facsimiles should be included in the research data environment. As a first step, main representatives of each lexicon type are represented. The plain text is marked up in TEI and is provided with metadata which include standardised form of the head word, information about its origin, foreign etymon for loan words (such as ἀπόστολος for апостолъ) and topic group designation. Word entries contain metadata about lexicographical or literary sources. In case that the facsimile is available, it will be presented along with the text form of the lexicon.

The project is aimed, on one hand, at the representation of the most important lexicons in marked up textual form and as facsimile (if possible). On other hand, the search engine embedded in the infrastructure will make it possible to create enquiries to the data and make selections such as: words according to their origin, literary or lexicographical sources, words belonging to a particular topic groups and so on. The manuscripts will be semantically exploited and concept-based, interlinked with knowledge resources in the linked open data framework (for more details see Kovalenko et al., 2016). For example, interlinking with Europeana will increase the accessibility and reusability of the data. Thus, the project will make Russian lexicons available and open them for further research and for public curiosity.

The poster will introduce the collaborative approach, the roles of the partners and their contributions, discuss the technical framework based on the research approach and present recent results the first time.

The project is connected to the COST action IS 1305 European Network of electronic Lexicography (ENeL).

Acknowledgements

This research was supported by the RGNF (the Russian Humanitarian Scientific Fund), project № 16-34-01008 "Russian manuscript dictionaries as a cultural phenomenon: the history of the ganre and literary context".

Bibliography

Kovalenko, K., Wandl-Vogt, E., Schopper, D., Declerck, T. (2016). "Opening up Russian Manuscript Lexicons for Cultural Heritage Studies." In *El'Manuscript–2016. Rašytinis palikimas ir informacinės technologijos. VI tarptautinė mokslinė konferencija. Pranešimai ir tezės. Vilnius, 2016 m. rugpjūčio 22–28 d.* Vilnius, pp. 71-74.

Kovalenko, K. (2016). "On the Classification of the Russian Manuscript Dictionaries." In *Words across History: Advances in Historical Lexicography and Lexicology.* Las Palmas de Gran Canaria, pp. 276-286.

Kovtun, L.S. (1975). *Leksikografiya v Moskovskoy Rusi XVI — nachala VXII v.* [*Lexicography in the Moscow Rus' in the 16th - Beginning of the 17th Century*]. Leningrad: Nauka.

Kovtun, L.S. (1989). *Azbukovniki XVI-XVII vv. (starshaya raznovidnost)* [*Azbukovniki of the 16th and 17th Centuries (the Earliest Edition)*]. Leningrad: Nauka.

Editing Melville's "Billy Budd" with TextLab, Juxta Editions, and MEL Catalog

Nick Laiacona
nick@performantsoftware.com
Performant Software Solutions, United States of America

Overview

During the creation of the Melville Electronic Library (MEL), we have developed a suite of software applications that, taken together, provide a powerful system for the creation of critical archives. In this software demonstration, we will demonstrate how these applications work, using the editing of Herman Melville's *Billy Budd* as an example. We will demonstrate the editorial process from the initial encounter with a manuscript leaf to the publishing of a web based reading text complete with a critical apparatus and editorial notes.

We will demonstrate the use of the following software applications:

- TextLab
- Juxta Editions
- MEL Catalog

Textlab

Billy Budd comes to us in the author's original manuscript, existing print editions, and many other media instantiations. This multiplicity is what John Bryant, the chief editor of MEL, refers to as a “fluid text.” (2002). TextLab aids in the transcription of the manuscript, the presentation of the manuscript facsimile, and the production of a base text. TextLab also aids in secondary editing, which is the process of identifying alternative readings at points of revision in the editing process. The ultimate aim is to retain the fluid quality of the text using digital tools.

TextLab’s team environment allows several editors to work simultaneously and independently on the hundreds of leaves of manuscript material. The interface pairs an XML editor with a high resolution image, and deep zoom functionality, allowing close-up viewing of the manuscript.

Textlab transforms TIFF images of the manuscript leaves into "pyramidal TIFFs" and serves them using a IIIF compliant image server.

The editor provides a set of XML elements and controlled vocabularies, which are pre-established by the editorial protocols of the edition. The transcriber can mark revision sites on the image using a drawing tool and then link these to the corresponding XML elements that record them. The transcriber can also validate the XML and preview the diplomatic rendering of the leaf before submitting it for inclusion in the edition.

As transcribers complete their work, the lead editor assembles the submitted transcriptions into a base text. TextLab aids in this process too, allowing the editor to reorder leaves and organize chapters without having to change the XML. In the end, a base text is produced, with diplomatic renderings of each leaf coordinated with the facsimile images. An XML version of the base text is stored in MEL Catalog, which is then retrievable by Juxta Editions.

Juxta Editions

Once a base text has been prepared using TextLab, it is then loaded into Juxta Editions twice. The first copy of the base text is kept for collation with other existing editions. Juxta creates a heat map visualization as well as a side-by-side comparison of the existing versions of the text. The second copy of the base text is then lightly edited for grammar and consistency. This produces a reading text which can then be added to the set of texts being collated.

The reading text is then annotated by the project team. As editors annotate the text, they can link passages to the MEL Catalog's database. Editors may link passages in the text itself or in the editorial notes.

When the text of "Billy Budd" is ready for publication to MEL, it is published using a customized Jekyll template. Visitors to the site will be able to read the text, trace Melville's words back to the original manuscript, and explore alternative readings through secondary editing and collation.

MEL Catalog

The MEL Catalog is a database of places, events, artwork, people, and texts from Melville's life and work. It is linked to name authorities such as the Getty Art and Architecture Thesaurus, so that the edition is compatible with linked open data. MEL Catalog can be used to search MEL as well as cross-walk on the objects it collects. MEL Catalog also has support for storing GIS data which can then be fed to mapping software the travel itineraries of persons both real and fictional.

Impact

Through a real-world example, this poster session will demonstrate how three specialized tools, used in conjunction, create a flexible, powerful and coherent process, which allows scholars to contend with source materials and produce digital editions more easily, and more quickly. We will also provide an explanation of how attendees can use these, and other compatible tools, for their own scholarship.

Bibliography

Bryant, J. (2002). *The Fluid Text.* Ann Arbor: U of Michigan, http://dx.doi.org/10.3998/mpub.12024

Generative Model For Latent Reasons For Modifications

David Lassner
davidlassner@mailbox.tu-berlin.de
TU Berlin, Germany

Problem

The idea that writing makes its way from the authors first draft manuscript to the intended reader without any detours or modifications is often inaccurate and oversimplified. In general, the author or a close person performs corrections and stylistic modifications in subsequent iterations. Additionally, there may be an editor or even an official censor who perform censorship of too private or too extreme parts of the document. The different versions of a document generated by these correction layers often become intransparent in printed versions of the document, while manuscripts are more likely to display traces of how the document has been modified to its current state. The digital scholarly edition "Letters and texts. Intellectual Berlin around 1800" (Baillot, 2016, IB in the following) combines genetic edition and entity annotation. The corpus encompasses literary and scholarly testimonies by a group of people, who influenced the intellectual Berlin between Enlightenment and Romanticism. The genetic encoding gives precise information regarding deletions and additions in the manuscript text. However, the reason for these modifications is not encoded. Three main domains for reasons why to modify such a document as a letter in the intellectual context of the time around 1800 have been identified:

1. Correction of mistakes
2. Stylistic modification
3. Moral censorship based on the topic

This paper proposes an unsupervised machine learning approach, which assigns the according reason to every modification. The proposed method focuses on dealing with stylistic modifications and moral censorships. I am aiming to increase the accessibility to manuscripts, by providing a structure for the modifications and to assist in evaluation of certain modifications. Furthermore the proposed method may be applied on different editorial problems, which I will discuss in the Outlook section.

Method

As brought up in the Problem section, the proposed method focuses on stylistic and moral censorship reasons, based on the assumption that these two types of reasons relate to the topic of the modification. I convey a generative topic model, that is based on Latent Dirichlet Allocation (D. Blei, Ng, & Jordan, 2003) and is able to take into account the structural information of modifications. There exists a wide range of topic models that customize LDA and many of these take into account additional structural information. To replace the Bag-of-words approach by introducing structural information about the word order is a major field of LDA research (Gruber, Rosen-Zvi, & Weiss, 2007; Wallach, 2006). Moreover there exists a lot of research on topic hierarchies (D. M. Blei, Griffiths, & Jordan, 2010; Paisley, Wang, Blei, & Jordan, 2015). LDA has also been modified to work with graph-structured documents (Xuan, Lu, Zhang, & Luo, 2015). However I am not aware of any literature that shows how to model modification reasons in a corpus of natural language.

Figure 1 illustrates the conceptual functioning of the method from left to right. As input on the left, a collection of documents is given. The documents have parts marked as modified. The generative model in the center infers reasons by taking into account all text, inside and outside the modifications. Every reason may stand for a stylistic, or a certain moral censorship reason (e.g. political, religious). On the right side, the model outputs a reason-modification assignment.

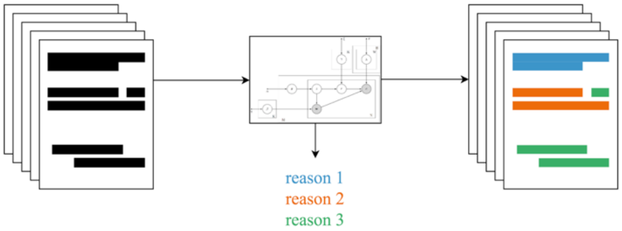


Figure 1: The generative model in the center receives input documents with modifications. It outputs reasons for modification and a reason assignment to each modification

In addition to the LDA latent variables, I introduce a topic-reason variable γ, a word-reason-modification tendency λ and a token-reason assignment r. The complete model in plate notation is shown in Figure 2. c (observed) models whether a token has been modified. For every topic, γ holds a distribution over reasons, which may cause a modification. For most modifications this distribution should be sparse, for example if a censor crosses out a sentence that discusses the financial situation of the author, the the topic and the reason for censorship would be identical. On the contrary a stylistic modification wouldn't always have one or two clear corresponding topics.

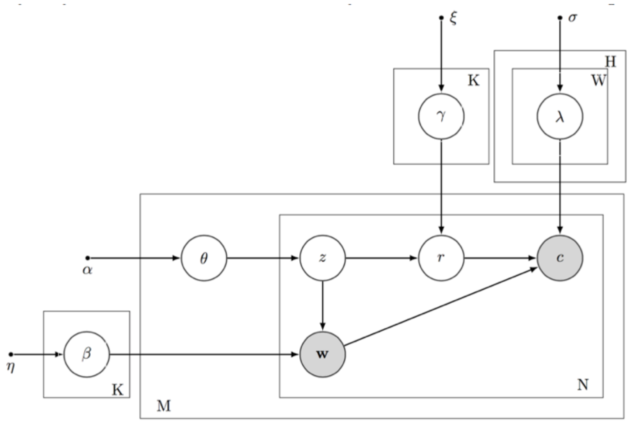


Figure 2: Plate notation of the model. The left four circled variables represent LDA, the right ones the modification part

For every token and every reason, λ holds a distribution over two states, namely whether the token tends to be modified for this reason. There may be token, that are representative for a topic, but they nonetheless do not tend to be modified. The categorical variable *r* represents the reason assignment at that position.

The latent variables can be iteratively approximated using Variational Inference (Bishop, 2006; Zhao, 2013).

Intermediate results

In this section, evaluation methods on toy data are discussed and characteristics of the IB data set, as well as preparation steps and first intermediate results are presented.

Toy data

To evaluate the characteristics of this method, experiments with artificial toy data can be performed. The generative model described above can be employed for inference as well as for generating artificial documents with modifications. A typical experiment to evaluate a generative model is conceived as follows:

1. Initialize the latent variables of the model randomly
2. Generate documents with modifications
3. Re-initialize the latent variables
4. Try to infer the latent variables from the generated documents

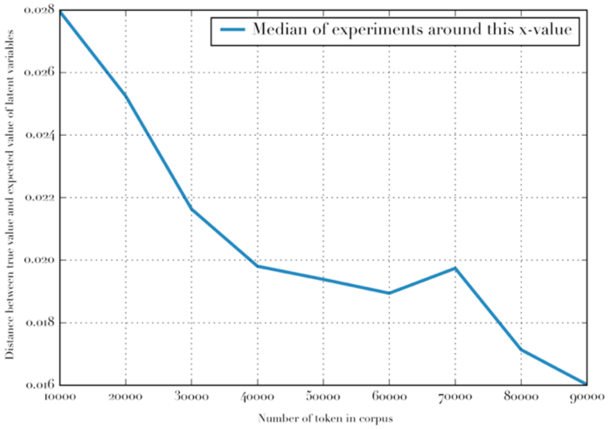


Figure 3: Experiment with 585 generated corpora varying the size from 10.000 to 90.000 token and performing inference for z, r, θ und γ, leaving the remaining model parameters fixed

I have performed a series of experiments to investigate the sensitivity of the model to corpus size. As expected, the accuracy of the model increases with an increasing amount of data. Figure 3 shows a decreasing distance between the true value and the expected one, when increasing the size of the data set. The accuracy does also largely depend on the sparsity of the concentration factors, which means that in order to predict the minimal size a real data set should have, one has to come up with according prior concentration factors for the Dirichlet variables.

IB data set

To apply this method on the IB data set, some preprocessing steps are necessary. Apart from standard natural language preprocessing, one has to filter out all corrections of mistakes.

Table 1 shows the change of data set characteristics that are caused by the pre-processing. A lot of modifications have been considered to be corrections of mistakes and thus have been filtered out.

The visualization in Figure 4 reveals a great variety in the structure of the modifications. The figure shows the state of all tokens of two letters from the IB data set. The upper letter contains a lot of small changes, where often a green (added part) and red (deleted part) occur as a combination. The letter below contains a lot of longer deleted parts, concluding, that the letter above contains corrections of mistakes, whereby the lower contains modifications related to the topic.

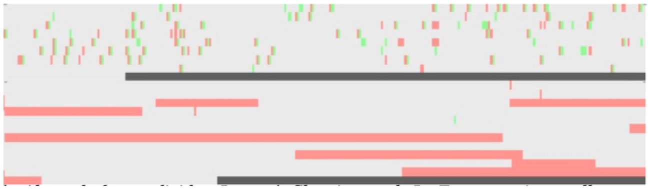


Figure 4: Above dark grey divider: Letter 4, Chamisso to de La Foye contains small corrections. Below: Letter 14, Dorothea Tieck to Uechtritz contains larger modifications. Deletions (red), additions (green)

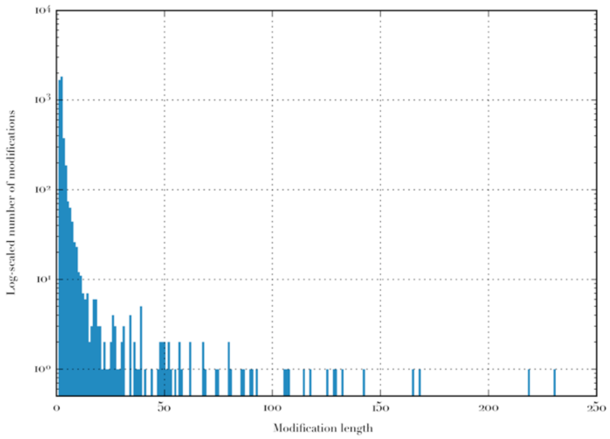


Figure 5: 93% of the modifications are shorter than 6 token. Two outliers of length 349 and 470 have not been included in the visualization

The size of the modifications seems to be a criterion to distinguish between corrections and topic related modifications. The distribution over the length of modifications however, reveals, that a lot of the modifications in the data set are small, thus likely to be corrections of mistakes (Figure 5).

The first preliminary experiments have been carried out with a binary setup: The model should distinguish between a) one particular moral censorship reason and b) everything else. To do so, prior knowledge about the topics has been introduced to the model in form of a keyword.

For example the topic sickness has been introduced to the model by the keyword “Krankheit”. The first results on this look very promising, as they reveal a precision = 1 and a recall = 0,67.

Outlook

In the near future, I will undertake further experiments with the IB data set. To do so, I will incrementally increase the number of modification reasons. The results will be made accessible as part of the IB corpus, making the permeability between editorial and algorithmic work more visible and accessible to all interested DH communities for reuse.

In a further step, I would like to look into different applications of this method. A promising idea would be, to look into different editions of the same text and consider each difference as a modification.

Bibliography

Baillot, A. (Ed.). (2016). *Letters and texts. Intellectual Berlin around 1800.* Berlin: Humboldt-Universität zu Berlin. Retrieved from http://www.berliner-intellektuelle.eu/. Please visit the web page for an up-to-date version.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer Science+Business Media, LLC.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). *The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. J.* ACM, 57(2), 1–30

Blei, D., Ng, A., & Jordan, M. (2003). *Latent Dirichlet allocation.* JMLR.

Gruber, A., Rosen-Zvi, M., & Weiss, Y. (2007). Hidden Topic Markov Models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. JMLR.

Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I. (2015). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 37(2).

Wallach, H. (2006). Topic Modeling: Beyond Bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning.* New York: ACM.

Xuan, J., Lu, J., Zhang, G., & Luo, X. (2015). Topic model for graph mining. *IEEE Transactions on Cybernetics*, 45(12).

Zhao, W. X. (2013). *Varitional Methods for Latent Dirichlet Allocation.* Retrieved from http://net.pku.edu.cn/~zhaoxin/vEMLDA.pdf (accessed on 1st of November 2016)

Design of a corpus and an interactive annotation tool for graphic literature

Jochen Laubrock
laubrock@uni-potsdam.de
University of Potsdam, Germany

Sven Hohenstein
hohenstein@uni-potsdam.de
University of Potsdam, Germany

Eike Martin Richter
eike.richter@uni-potsdam.de
University of Potsdam, Germany

The project "Hybrid Narrativity" combines work by language and literature studies, cognitive psychology, and computer science with the overarching goal to arrive at an empirically founded narratology of graphic literature, including comics and graphic novels. Comics and graphic novels provide a unique cultural form that has developed its own vocabulary, allowing for a fascinating interplay of text and visual art. After a period of neglect, they have recently been theoretically analyzed in detail by scholars in the arts, humanities and linguistics (McCloud, 1993; Groensteen, 2007; Cohn, 2013). Our aim is to provide an empirical testbed for these theories. The foundation of this endeavor is a large collection of graphic novels, which are annotated using a variety of methods. These include high-level descriptions of the work, mid-level descriptions of pages and panels, including the actors / characters, text, objects, and panel transitions, and low-level descriptions of visual elements in terms of descriptors developed in computer vision such as color histograms, GIST, SIFT, and SURF features. We are currently evaluating the addition of mid-level features from deep networks trained on photographs of real-world scenes, with quite promising first results.

These descriptions in terms of material properties are complemented by eye-tracking data, providing an empirical measure of the reader-level attention distribution and the time course of attention shifts. Thus, the digitized representation of literary and artistic works includes information on the side of "recipients", that is, readers, viewers, spectators and appreciators with their psychological and physiological responses. In a first step, eye-tracking data on a small sample of pages from selected works were collected from a large number of participants in order to evaluate general principles of attentional selection in graphic literature. Results show that reading of graphic novels is primarily governed by reading of text, and that inspection of graphical elements is apparently governed by top-down selection of story-relevant elements. In perspective, eye-tracking data will be collected for each of the works in the corpus, using a sample of pages and a smaller number of readers.

A graphical annotation tool is in development and has first been released to the public at DH 2016. This tool is based on an XML dialect that allows for the annotation of language as well as graphical elements. Future versions will include OCR support for comics fonts, and provide customizable annotation schemes, allowing other researchers to implement their own research ideas. We will also briefly present ideas on the potential to incorporate gaze-based interaction in the user interface of the tool, e.g., for the intuitive selection of objects, which will become important with the projected availability of low-cost eye trackers in the near future.

We have developed the Graphic Novel Markup Language (GBML) as an extension of John Walsh's Comic Book Markup Language (CBML; Walsh, 2012) to facilitate the description of graphical elements. These descriptions are imperative for defining regions-of-interest based mapping of eye movement data to the stimulus material. Material has been annotated using our editor, and a custom R package is under development and in use for statistical analysis of eye movements. Visual features are currently extracted using OpenCV (Bradski, 2000, 2016) and VLFEAT (Vedaldi & Fulkerson, 2008) libraries from Python and Matlab, since R does not yet provide sufficiently extensive packages for this purpose (for a promising approach see imager, Barthelmé, 2016). Deep features are based on Deep Gaze II (Kümmerer, Wallis, & Bethge, 2016), which is in turn based on the VGG-19 network (Simonyan & Zisserman, 2014). A description of artworks in terms of visual features has shown promising results in other domains (Elgammal & Saleh, 2015).

A number of analyses using the corpus data show the potential for comparative studies as well as detailed study of individual works. Many of the testable hypotheses can be derived from the theoretical work cited above. For example, McCloud (1993) speculated that the cognitive effort of the recipient depends on the kind of panel transition, or that the empty space between panels is used to signal the passage of time. We provide empirical support for both of these hypotheses. Other examples include visual trends that can be identified across time or between regions, linguistic and visual analyses that can be used to compare text and visual complexity between different genres, and network analyses of interactions between characters that allow for an easy quantification and visualization of roles within a work, and for a comparison between works. They can also be used, e.g., to compare the complexity of a novel and its adaptation to the graphic novel format.

An in-depth analysis of a single work, the graphic novel adaptation of Paul Auster's City of Glass by Paul Karasik and David Mazzucchelli, shows that readers of graphic literature benefit from a specific expertise in decoding the different channels of information conveyed by image and text. Comics experts spend significantly more time on the image part of the panels, and this is correlated with a significantly deeper understanding of the narrative. New data suggests that this pattern replicates across samples, labs, and languages.

Taken together, we present the design of a corpus of graphic literature that is annotated using a variety of levels, including readers' eye movements. Ideas for how to make use of these data for interactive future versions are developed, and analyses of the collected data in terms of description as well as reception of the works of art are presented.

Bibliography

Barthelmé, S. (2016). imgager: Image Processing Library Based on 'CImg'. R package version 0.31. [Computer Software: https://CRAN.R-project.org/package=imager]

Bradski, G. (2000). The OpenCV library. Dr. Dobb's Journal of Software Tools, 25(11):120, 122–125.

Cohn, N. (2013). The Visual Language of Comics. Introduction to the Structure and Cognition of Sequential Images. London: Bloomsbury.

Elgammal, A. & Saleh, B. (2015). Quantifying Creativity in Art Networks. 6th International Conference on Computational Creativity (ICCC'15).

Groensteen, T. (2007). The System of Comics. Translated by B. Beaty and N. Nguyen. Jackson, MI: University of Mississippi Press.

Kümmerer, M., Wallis, T.S.A., & Bethge, M (n.d.).: DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv:1610.01563

McCloud, S. (1993). Understanding Comics: The Invisible Art. New York: Harper Collins.

Simonyan, K. & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In: CoRR abs/1409.1556. url: http://arxiv.org/abs/1409.1556.

Vedaldi, A. & Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. [Computer Software: http://www.vlfeat.org/]

Walsh, J. (2012). Comic Book Markup Language: An Introduction and Rationale. Digital Humanities Quarterly, 6(1).

SLaTE: A System for Labeling Topics with Entities

Anne Lauscher
anne@informatik.uni-mannheim.de
University of Mannheim, Germany

Federico Nanni
federico@informatik.uni-mannheim.de
University of Mannheim, Germany

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de
University of Mannheim, Germany

In recent years, the Latent Dirichlet allocation (LDA) topic model (Blei, Ng, and Jordan, 2003) has become one of the most employed text mining techniques (Meeks and Weingart 2012) in the digital humanities (DH). Scholars have often noted its potential for text exploration and distant reading analyses, even when it is well known that its results are difficult to interpret (Chang et al, 2009) and to evaluate (Wallach et al, 2009).

At last year's edition of the Digital Humanities conference, we introduced a new corpus exploration method able to produce topics that are easier to interpret and evaluate than standard LDA topic models (Nanni and Ruiz, 2016). We did so by combining two existing techniques, namely Entity linking and Labeled LDA (L-LDA). At its heart, our method first identifies a collection of descriptive labels for the topics of arbitrary documents from a corpus, as provided from the vocabulary of entities found within wide-coverage knowledge resources (e.g., Wikipedia, DBpedia). Then it generates a specific topic for each label. Having a direct relation between topics and labels makes interpretation easier, and using a disambiguated knowledge resource as background knowledge limits label ambiguity. As our topics are described with a limited number of unambiguous labels, they promote interpretability, and this may sustain the use of the results as quantitative evidence in humanities research (Lauscher et al, 2016).

The contributions of this poster cover the release of: a) a complete implementation of the processing pipeline for our entity-based LDA approach; b) a three-step evaluation platform that enables its extensive quantitative analysis.

Entity-based Topic Modeling Pipeline

Figure 1 illustrates the computational pipeline of our system; python classes are represented in rectangles. First of all, a set of text files is imported into the system and several preprocessing steps are applied to the textual content. Next, the data is sent to the entity linking system TagMe (Ferragina and Scaiella, 2010), which disambiguates against Wikipedia. As a result of this step, for each document a set of related Wikipedia entities is retrieved. Now, the data is inserted into a MySQL database.

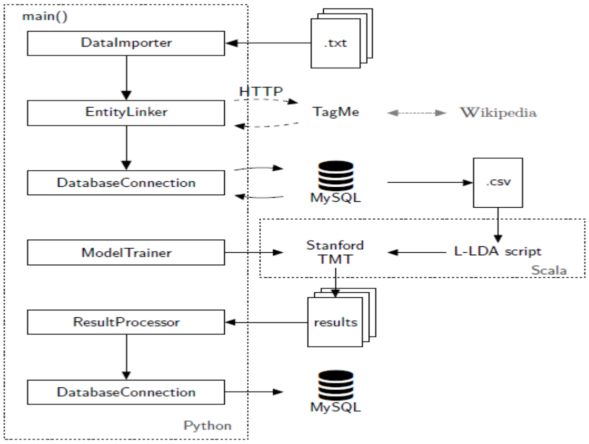


Figure 1. Architecture of the pipeline

Afterwards, the TF-IDF measure is computed over the entities, which we use to rank all the entities for each document in descending order. Then, the top k entities as well as their corresponding documents are exported into a comma-separated values file that is given as input to the L-LDA implementation of the Stanford Topic Modeling Toolbox. Finally, after running L-LDA and applying several post-processing steps, we obtain a document-topic distribution saved in the database in which each topic is described by an unambiguous label linked to Wikipedia.

The whole source code is available for public download on Github. Given a working Python, Java, and Scala runtime as well as a running MySQL installation our pipeline is ready directly out-of-the-box. The specific configuration according to the user's needs can be made via a simple text file.

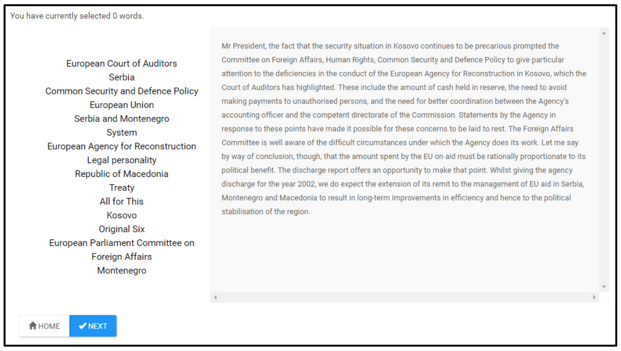
Three-Step Evaluation Platform

Document Labels

In order to assess the quality of the detected entities as labels we developed a specific browser-based evaluation platform, which permits manual annotations. This platform presents a document on the right of the screen and a set of possible labels on the left (See Figure 2). Annotators are asked to pick labels that precisely describe the content of each document. In case the annotator does not select any label, this is also recorded by our evaluation system.

Entities and Topic Words

In order to establish if the selected entities were the right labels for the topics produced, we developed two additional evaluation steps. Inspired by the topic intrusion task (Chang et al, 2009), we designed a platform that permits to evaluate the relations between labels and topics using two evaluation modes: For one evaluation mode (that we called Label Mode - Figure 3), the annotator is asked to choose, when possible, the correct list of topic-words given a label. For the other, he/she was asked to pick the right label given a list of topic words (aee Figure 4). In both cases, the annotator is shown three options: one of them is the correct match, while the other two (be they words or labels) come from other topics related to the same document.

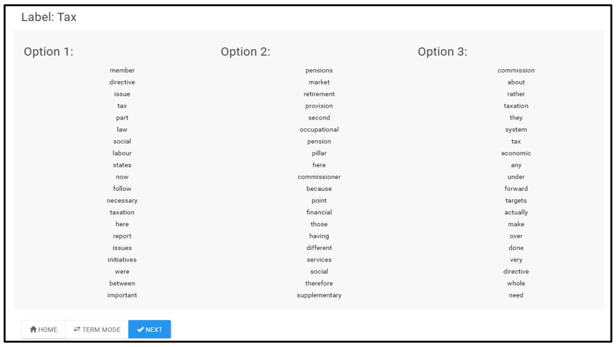


You have currently selected 0 words.

European Court of Auditors
Serbia
Common Security and Defence Policy
European Union
Serbia and Montenegro
System
European Agency for Reconstruction
Legal personality
Republic of Macedonia
Treaty
All for This
Kosovo
Original Six
European Parliament Committee on Foreign Affairs
Montenegro

Mr President, the fact that the security situation in Kosovo continues to be precarious prompted the Committee on Foreign Affairs, Human Rights, Common Security and Defence Policy to give particular attention to the deficiencies in the conduct of the European Agency for Reconstruction in Kosovo, which the Court of Auditors has highlighted. These include the amount of cash held in reserve, the need to avoid making payments to unauthorised persons, and the need for better coordination between the Agency's accounting officer and the competent directorate of the Commission. Statements by the Agency in response to these points have made it possible for these concerns to be laid to rest. The Foreign Affairs Committee is well aware of the difficult circumstances under which the Agency does its work. Let me say by way of conclusion, though, that the amount spent by the EU on aid must be rationally proportionate to its political benefit. The discharge report offers an opportunity to make that point. Whilst giving the agency discharge for the year 2002, we do expect the extension of its remit to the management of EU aid in Serbia, Montenegro and Macedonia to result in long-term improvements in efficiency and hence to the political stabilisation of the region.

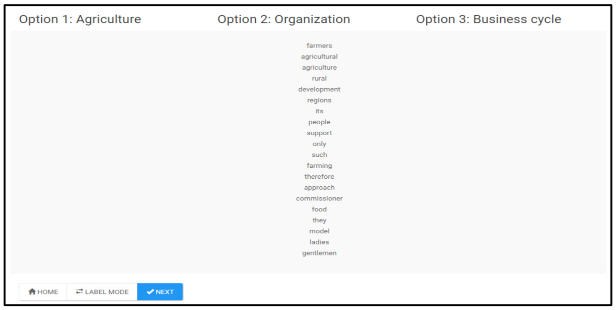
Figure 2. Entities as Labels evaluation interface.



Label: Tax

Option 1:	Option 2:	Option 3:
member	pensions	commission
directive	market	about
issue	retirement	rather
tax	provision	taxation
part	second	they
law	occupational	system
social	pension	tax
labour	pillar	economic
states	here	any
now	commissioner	under
follow	because	forward
necessary	point	targets
taxation	financial	actually
here	those	make
report	having	over
issues	different	done
initiatives	services	very
were	social	directive
between	therefore	whole
important	supplementary	need

Figure 3. Label-Mode Evaluation



Option 1: Agriculture Option 2: Organization Option 3: Business cycle

farmers
agricultural
agriculture
rural
development
regions
its
people
support
only
such
farming
therefore
approach
commissioner
food
they
model
ladies
gentlemen

Figure 4. Term-Mode Evaluation

Bibliography

Blei, D. M., Ng, A.Y., and Jordan, M. (2003) "Latent dirichlet allocation." *Journal of machine Learning research*.

Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C. and Blei, D.M. (2009) "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems*. 21. 288-296.

Ferragina, P., and Scaiella, U. (2010) "TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities)." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*.

Lauscher, A., Nanni, F., Ruiz Fabo, P., and Paolo Ponzetto, S. (2016). "Entities as topic labels: combining entity linking and labeled LDA to improve topic interpretability and evaluability." IJCol-Italian journal of computational linguistics. 2(2): 67-88.

Meeks, E., and Weingart, S.B. (2012). "The digital humanities contribution to topic modeling." *Journal of Digital Humanities* 2.1.

Nanni, F., and Ruiz Fabo, P. (2016) "Entities as topic labels: Improving topic interpretability and evaluability combining Entity Linking and Labeled LDA." *DH2016*.

Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009). "Evaluation methods for topic models." *Proceedings of the 26th Annual International Conference on Machine Learning*.

Journaux intimes francophones en Russie au XIXe siècle : une approche textométrique de l'interaction entre langues et cultures

Alexei Lavrentiev
alexei.lavrentev@ens-lyon.fr
IRHIM, CNRS & ENS de Lyon, France

Michèle Debrenne
micheledebrenne@gmail.com
Université d'État de Novossibirsk, Russie

Dmitry Dolgushin
d_dolgushin@mail.ru
Université d'État de Novossibirsk, Russie

Nina Panina
pa.nina@mail.ru
Université d'État de Novossibirsk, Russie

Andrey Borodiknin
borodichin@spsl.nsc.ru
Branche Sibérienne de l'Académie des sciences, Russe

Serge Heiden
slh@ens-lyon.fr
IRHIM, CNRS & ENS de Lyon, France

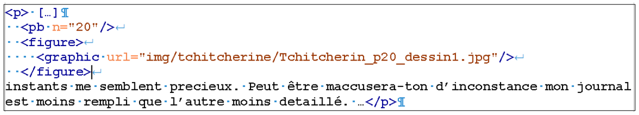
La rédaction de journaux intimes en français était un phénomène assez répandu parmi les élites russes au XIX^{e} siècle. Les aristocrates de l'époque apprenaient souvent le français dès le plus jeune âge et continuaient à pratiquer cette langue tout au long de leur vie. Le français n'était cependant pas leur langue maternelle, et leurs écrits offrent des données précieuses sur l'interaction entre les deux langues et cultures. Par ailleurs, certains journaux présentent de grandes qualités littéraires et artistiques (grâce aux dessins d'auteur qui les accompagnent) et méritent d'être plus largement connus par la communauté académique et le grand public.

Le projet présenté dans ce poster est porté par une équipe franco-russe de l'Université d'état de Novossibirsk et de l'École normale supérieure de Lyon. Il vise à analyser et à publier numériquement plusieurs corpus, dont les journaux d'Alexandre Tchitchérine et d'Olga Orlova-Davydova.

A. Tchitchérine, lieutenant au régiment Semionovski de l'armée russe, a tenu son journal de septembre 1812 à août 1813, son manuscrit contient 83 illustrations, dont la plupart en couleurs. Son texte a été traduit en russe et publié (Chicherin 1966), mais cette traduction dont les qualités littéraires sont indéniables, est parfois inexacte et présente des lacunes. L'original français de ce journal est inédit.

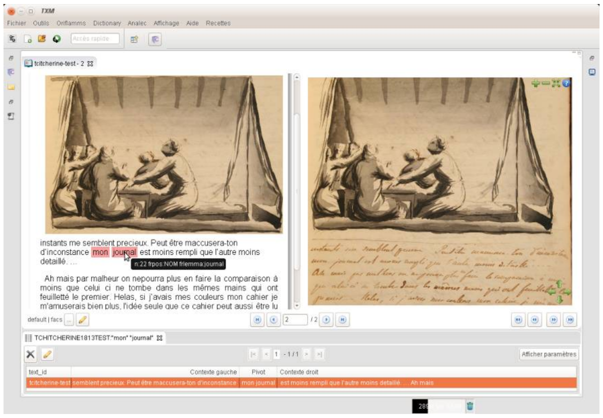
Les journaux d'O. Orlova-Davydova (1814–1876) sont le fruit d'un vaste projet autobiographique que l'auteure a mené tout au long de sa vie. Ces journaux n'ont jamais fait l'objet d'une recherche scientifique. Dans le cadre du présent projet les passages francophones de ses journaux (1830-1847) servent de corpus pour étudier le bilinguisme de l'auteure et pour déceler les voies de formation de personnalités bi- ou plurilingues dans la Russie du XIX^{e} siècle (Debrenne 2016a et 2016b). Le noyau du corpus est constitué par une copie des journaux d'Orlova-Davydova réalisée vraisemblablement à sa demande et corrigée par sa propre main. Ce document est conservé à Novossibirsk, à la Bibliothèque scientifique et technique de la Branche Sibérienne de l'Académie des sciences russe (BST). Il fait partie du fonds patrimonial Tikhomirov (http://www.spsl.nsc.ru/rbook/ogl_tix.html).

Les journaux du corpus sont transcrits et « encodés par stylage » dans un logiciel de traitement de texte (Microsoft Word) de façon à permettre une conversion quasi-automatique dans un format XML-TEI enrichi d'un certain nombre d'annotations. Il s'agit de repérer et de classer les différents types d'erreurs par rapport au français standard de l'époque (orthographiques, grammaticales ou lexicales), de recenser les différents types d'entités nommées présents (index hiérarchique), ainsi que les différents passages rédigés en russe. La conversion se fait en deux étapes : d'abord à l'aide de l'outil de conversion en ligne Odette (Glorieux 2015), puis avec une feuille de transformation XSLT spécifique au projet. Après validation formelle, la forme finale des sources est importée dans la plate-forme TXM (Heiden et al. 2010). TXM permet d'analyser et de publier des corpus complexes encodés en XML-TEI en combinant des outils de traitement automatique de la langue (TAL :tokenisation paramétrable, étiquetage morphosyntaxique avec TreeTagger), de philologie numérique (éditions présentant de façon synoptique le fac-simile de la source et plusieurs niveaux de transcription), et d'analyse qualitative (recherche de motifs lexicaux avec le moteur plein texte CQP) et quantitative (analyse de spécificités, analyse factorielle de correspondances, etc.). Le projet utilisera les outils de TXM pour lire de façon intégrée l'édition du texte avec les illustrations correspondantes. Il permettra également de synthétiser le vocabulaire utilisé juste devant les passages en russe et étudier les emprunts français propres à chaque auteur par rapport à la langue russe. Les illustrations ci-dessous présentent un exemple d'encodage XML-TEI dans les sources d'un dessin situé au début d'une page du journal de Tchitchérine (Fig. 1) suivi de la copie d'écran d'une visualisation synoptique correspondante dans TXM de la transcription comprenant le dessin et le fac-similé du manuscrit (Fig. 2).



```
<p>·[…]¶
··<pb·n="20"/>↵
··<figure>↵
····<graphic·url="img/tchitcherine/Tchitcherin_p20_dessin1.jpg"/>↵
··</figure>↵
instants·me·semblent·precieux.·Peut·être·maccusera-ton·d'inconstance·mon·journal
est·moins·rempli·que·l'autre·moins·detaillé.·…</p>¶
```

Fig. 1. Encodage XML-TEI d'une illustration insérée dans le corps du texte


TXM
Fichier Outils Oriflamms Dictionary Analec Affichage Aide Recettes
Accès rapide
tcitcherine-test - 2
n:22 frpos:NOM frlemma:journal
default | facs
2 / 2
TCHITCHERINE1813TEST:"mon" "journal"
1 - 1 / 1
Afficher paramètres

instants me semblent precieux. Peut être maccusera-ton d'inconstance mon journal est moins rempli que l'autre moins detaillé. ...

Ah mais par malheur on nepourra plus en faire la comparaison à moins que celui ci ne tombe dans les mêmes mains qui ont feuilletté le premier. Helas, si j'avais mes couleurs mon cahier je m'amuserais bien plus, l'idée seule que ce cahier peut aussi être lu

text_id	Contexte gauche	Pivot	Contexte droit
tcitcherine-test	semblent precieux. Peut être maccusera-ton d'inconstance	mon journal	est moins rempli que l'autre moins detaillé. ... Ah mais

Fig. 2. Affichage synoptique d'une transcription et d'un facsimilé d'un texte doté d'une illustration sous TXM en mode « retour au texte » depuis une concordance (de l'expression « mon journal » dont l'occurrence est mise en évidence dans la transcription)

Remerciements

Ce travail a bénéficié du soutien la Fondation scientifique de Russie pour les sciences humaines (RGNF) [16-24-08001] et de la Fondation Maisons des Sciences de l'Homme (FMSH).

Note

Tous les corpus édités dans le cadre du projet seront publiés en ligne via un portail TXM et sur le site de la BST (bibliothèque numérique du fonds Tikhomirov). Des exemples de réalisations similaires peuvent être consultés sur le portail de démonstration.

Bibliography

Chicherin, A. V. (1966). *Dnevnik Aleksandra Chicherina*. Moscou: Nauka.

Debrenne, M. (2016a). "Comparative error-based analysis of the copied Davydova's diary and its original." *Vestnik NGU. Seria Lingvistika i mezhkulturnaia kommunikacia* 14(3): 59-74.

Debrenne, M. (2016b). "The French Language in the Diaries of Olga Davydova. An example of Russian-French Aristocratic Bilingualism." In van Strien-Chardonneau, S. and Kok Escalle, M.-C. (eds), *Le français, langue de l'intime à l'époque moderne et contemporaine*. Amsterdam: Amsterdam University Press B.V., pp. 125-142.

Glorieux, F. (2015). *Le traitement de textes (odt) pour produire des documents structurés (XML/TEI) – Odette*, http://resultats.hypotheses.org/267.

Heiden, S., Magué, J-P. and Pincemin, B. (2010). "TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement." In Sergio Bolasco, I. C. (ed), *Proceedings of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*. Rome, Edizioni Universitarie di Lettere Economia Diritto, vol. 2, pp. 1021-1032. <halshs-00549779>

Archivlo: Digitizing the Archival Research Workflow

Zoe LeBlanc
zoe.leblanc@vanderbilt.edu
Vanderbilt University, United States of America

E. Kyle Romero
eulogio.k.romero@vanderbilt.edu
Vanderbilt University, United States of America

This poster describes the development and uses of Archivlo, an application for improving the archival research workflow and enabling a more collaborative digital research community. In recent years, digital history has emerged as a vibrant subfield of the digital humanities community (see Robertson, 2016; and Weingart, 2016).Currently, the majority of digital history projects rely on digitized corpuses or community compiled datasets. However, the archival materials used in these projects represent only a small fraction of the archival sources that scholars currently utilize in their research.

Moreover, the proliferation of digital cameras and scanners has resulted in a wealth of archival material for scholars, but this digitized archival data is usually scattered across hard drives. To organize this data, scholars currently either keep notes or re-purpose bibliographic software.

Data management software provide some solutions to dealing with this abundance of material (such as Devonthink, Evernote, Zotero, and most recently, Tropy) but individual scholars often must invest a great deal of energy and time replicating the organizational structure of the archives to make sense of their research. This siloed approach to archival research makes finding information about archival collections or other scholars working in the archives difficult. Archivlo is designed to solve these problems, and create a more coherent workflow for organizing archival data.

This poster will outline the development and design of Archivlo, from the early idea stages to our initial beta model. Archivlo is currently in progress, and the poster will share our experience building a web-based application, as well as designing a user interface that privileges data interoperability and flexibility. Archivlo is written in Python and Angular, and is fully open-source on Github. To access archival data, Archivlo utilizes archives' APIs and web page annotations to allow researchers to find collections. Users are able to save their archival collection research in their profile, and indicate whether they have worked in these archives or are interested in using the archive. This functionality adds efficiencies to how scholars locate and keep track of their archival research. Users can also export their records to multiple file formats, as well as other data management software, such as Zotero andDevonthink.

Additionally, Archivlo enables users to share their lists of visited and interested in archives, which we believe will help scholars share information about archives and potentially even form collaborations. For archivists, Archivlo can also provide data on user interest vis-a-vis usage of their archival collections. We believe our experience with Archivlo will be of interest to other digital humanities developers and project managers, as well as digital humanists who work with archival collections.

Previous efforts to encourage digital collaboration among researchers in archives have, with a few exceptions, largely faltered, with most of these projects requiring a high technical literacy to contribute to a database or extensive time to transcribe records (see Mostern and Arksey, 2016. Moreover, these efforts to construct large databases of archival data have been forced, through copyright restrictions, to limit their scope to material that is either from prior to the early twentieth century or born digital materials). Instead of requiring large resources to digitize materials or standardize collections, Archivlo presents an alternative solution to this problem - focusing on how scholars work with archives to enable more digital and collaborative research. We believe Archivlo will encourage more productive data management practices among scholars, and reduce inefficiencies in the archival research workflow. Much of Archivlo's goals remain experimental, and the opportunity to present our work at DH 2017 would help us share our progress and consider future directions for the tool.

Ultimately, we hope that Archivlo can help further the digital humanities ethos of digital collaboration, and present one solution for using tools to help foster digital research communities.

Bibliography

Mostern, R., and Arksey, M. (2016) “Don’t Just Build It, They Probably Won’t Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences”, *International Journal of Humanities and Arts Computing*, Volume 10 Issue 2, 205-224.

Robertson, S. (2016) “The Differences between Digital Humanities and Digital History” in *Debates in Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein, (University of Minnesota Press, 2016), 289-307

Weingart, S. (2016) “Acceptances to DH2016 (pt. 1)”, March 22, 2016, <https://scottbot.net/acceptances-to-dh2016-pt-1/>.

Supplementing Melody, Lyrics, and Acoustic Information to the McGill Billboard Database

Hubert Léveillé Gauvin
leveillegauvin.1@osu.edu
Ohio State University, United States of America

Nathaniel Condit-Schultz
natsguitar@gmail.com
McGill University, Canada

Claire Arthur
claire.arthur81@gmail.com
McGill University, Canada

Research Subject and Issues

Musical organization in popular music is a complex research subject. While significant work has been done in the last few decades to theorize popular music (e.g. Moore, 1992, 1995; Tagg, 2009; Biamonte, 2010; Attas, 2011), these theories tend to rely primarily on score-based parameters such as harmony and, to a lesser extent, melody and rhythm. However, in popular music, scholars have observed that audio-based features such as perceived loudness and timbre play an important role in musical organization (e.g., Everett, 2001; Temperley, 2007). To empirically evaluate the relevance of these features to various theories, it would be appropriate to supplement the traditional melodic and syntactic approaches with acoustical approaches. Empirically oriented musicological surveys of popular music have become more and more popular. However, as is the case with more traditional musicological research, studies on harmony still occupy the core of the research that has been published in the last decade (e.g. Mauch et al., 2007; de Clercq and Temperley, 2011, Temperley and de Clercq, 2013; Burgoyne et al., 2011; Léveillé Gauvin, 2015). This underrepresentation of large-scale research focusing on other parameters such as pitch content and perceived loudness in popular music may in part be attributed to the lack of corpora with such data.

Objectives

The aim of the current project is to gain unique insights into popular music by assembling a new database that will combine score-based and audio-based parameters. As such, we are releasing melodic and lyric transcriptions, as well select signal-processing data to supplement a subset of 100 songs from the already publicly available McGill Billboard Database (Burgoyne et al., 2011).

Sample

The McGill Billboard Database is a systematically sampled, professionally curated collection of harmonic transcriptions for more than 700 distinct songs that made the Billboard Hot 100 weekly charts between 1958 and 1991. Along with time-aligned chord transcriptions, each file features metadata regarding the title of the song, the performing artist, the chart date, the highest rank the song ever achieved on the Billboard Hot 100 chart, and the number of weeks the song spent on the charts. Our corpus focuses on a 100-song collection taken from the original McGill Billboard Database. More than 70 unique artists are represented, ranging from 1958 to 1991.

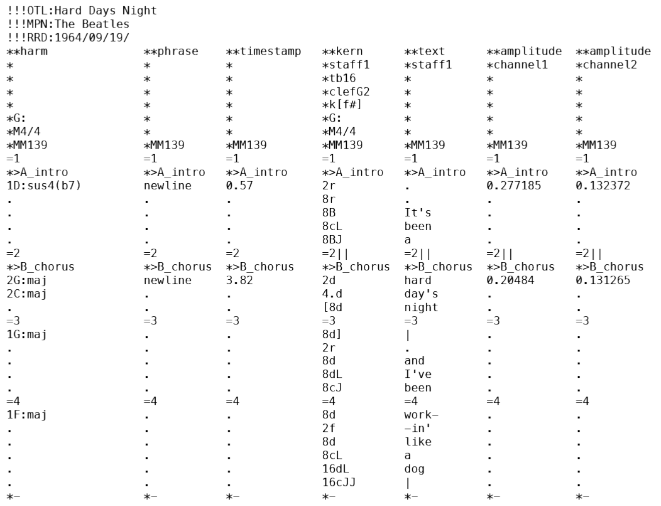
Transcription Process

All songs were divided and assigned randomly among a group transcribers, including the authors. Each transcriber was in charge of finding the appropriate recording matching the original chord transcription from the McGill Billboard Database and transcribing melodic and lyrical information using their preferred notational software. In order to alleviate discrepancies among the different transcribers, a set of guidelines in the form of a “Transcription Style Guide” was established and distributed prior to the transcribing process. This file stipulated minimum requirements for every transcription, as well as general instructions on how to notate potentially more challenging pitch nuances, such as slides, scoops, and ornaments.

We used the timestamps already available in the original McGill Billboard Database to automatically retrieve acoustical information. The data was encoded separately for the left and right channels, thus maintaining information related to stereo panning.

Encoding Format

We opted to encode the new transcriptions in the Humdrum format. Humdrum (Huron, 1995) is both a syntax to encode music information in ASCII representation and a set of tools dedicated to the manipulation of such files, alleviating the problem of having to write a dedicated parser. Figure 1 represents a typical Humdrum file in our corpus.



```
!!!OTL:Hard Days Night
!!!MPN:The Beatles
!!!RRD:1964/09/19/
**harm              **phrase    **timestamp   **kern      **text      **amplitude  **amplitude
*                   *           *             *staff1     *staff1     *channel1    *channel2
*                   *           *             *tb16       *           *            *
*                   *           *             *clefG2     *           *            *
*                   *           *             *k[f#]      *           *            *
*G:                 *           *             *G:         *           *            *
*M4/4               *           *             *M4/4       *           *            *
*MM139              *MM139      *MM139        *MM139      *MM139      *MM139       *MM139
=1                  =1          =1            =1          =1          =1           =1
*>A_intro           *>A_intro   *>A_intro     *>A_intro   *>A_intro   *>A_intro    *>A_intro
1D:sus4(b7)         newline     0.57          2r          .           0.277185     0.132372
.                   .           .             8r          .           .            .
.                   .           .             8B          It's        .            .
.                   .           .             8cL         been        .            .
.                   .           .             8BJ         a           .            .
=2                  =2          =2            =2||        =2||        =2||         =2||
*>B_chorus          *>B_chorus  *>B_chorus    *>B_chorus  *>B_chorus  *>B_chorus   *>B_chorus
2G:maj              newline     3.82          2d          hard        0.20484      0.131265
2C:maj              .           .             4.d         day's       .            .
.                   .           .             [8d         night       .            .
=3                  =3          =3            =3          =3          =3           =3
1G:maj              .           .             8d]         |           .            .
.                   .           .             2r          .           .            .
.                   .           .             8d          and         .            .
.                   .           .             8dL         I've        .            .
.                   .           .             8cJ         been        .            .
=4                  =4          =4            =4          =4          =4           =4
1F:maj              .           .             8d          work-       .            .
.                   .           .             2f          -in'        .            .
.                   .           .             8d          like        .            .
.                   .           .             8cL         a           .            .
.                   .           .             16dL        dog         .            .
.                   .           .             16cJJ       |           .            .
*-                  *-          *-            *-          *-          *-           *-
```

Figure 1: Example of a complete transcription in the Humdrum format

Impact and Future Work

We hope that this new collection of musically-rich data will yield new and unexpected research on popular music, and allow possibilities that were, up until now, virtually impossible. We believe that supplementing traditional score-based data (e.g. harmony and melody) with lyrics and loudness descriptors is a necessary step into developing a holistic theory of form in popular music.

Our plans for the future are manifold. We hope to increase the size of our corpus, with the goal to eventually provide complete annotations for every harmonic transcription in the McGill Billboard Database. We also wish to continue supplementing this corpus over the next several years with more detailed data. More specifically, we hope to have instrumental solos, drumming patterns, back vocals, and more acoustical information, including spectral annotations.

Acknowledgment

This research project was supported by the Fonds de recherche du Québec - Société et culture. Special thanks go to Dana DeVlieger, Lissa Reed and Gary Yim for their help.

Bibliography

Attas, R. (2011). "Sarah setting the terms: Defining phrase in popular music." Music Theory Online, 17(3).

Biamonte, N. (2010). "Triadic modal and pentatonic patterns in rock music." Music Theory Spectrum, 32(2): 95–110.

Burgoyne, J. A., Wild, J., and Fujinaga, I. (2011). "An expert ground-truth set for audio chord recognition and music analysis." Proceedings of the 12th International Conference on Music Information Retrieval. Mimi, FL, pp. 633–638.

De Clercq, T., and Temperley, D. (2011). "A corpus analysis of rock harmony." Popular Music, 30(1): 47–70.

Everett, W. (2001). The Beatles as Musicians: The Quarry Men through Rubber Soul. Oxford University Press.

Huron, D. (1995). The Humdrum toolkit: Reference manual. Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Léveillé Gauvin, H. (2015). "'The Times They Were A-Changin'': A database-driven approach to the evolution of harmonic syntax in popular music from the 1960s." Empirical Musicology Review, 10(3): 215–238.

Mauch, M., Dixon, S., and Harte, C. (2007). "Discovering chord idioms through Beatles and Real Book songs." Proceedings of the 8th International Conference on Music Information Retrieval. Vienna, Austria, pp. 255–258.

Moore, A. (1992). "Patterns of harmony." Popular Music, 11(1): 73–106.

Moore, A. (1995). "The so-called 'flattened seventh' in rock." Popular Music, 14(2): 185–201.

Tagg, P. (2009). Everyday Tonality: Towards a Tonal Theory of What Most People Hear. New York and Montreal: Mass Media Scholar's Press.

Temperley, D. (2007). "The melodic-harmonic 'Divorce' in rock." Popular Music, 26(2): 323–342.

Temperley, D., and De Clercq, T. (2013). "Statistical analysis of harmony and melody in rock music." Journal of New Music Research, 2(3), 187–204.

Exploring *Word Formation Latin*

Eleonora Litta
eleonoramaria.litta@unicatt.it
CIRCSE Research Centre
Università Cattolica del Sacro Cuore, Italy

Marco Passarotti
marco.passarotti@unicatt.it
CIRCSE Research Centre,
Università Cattolica del Sacro Cuore, Italy

Chris Culy
chrisculy@mac.com
Consultant @ CIRCSE Research Center

Paolo Ruffolo
CIRCSE Research Centre
Università Cattolica del Sacro Cuore, Italy

Introduction

Word Formation Latin (WFL) is a derivational morphology resource for Classical Latin, where lemmas are analysed into their formative components, and relationships between them are established on the basis of Word Formation Rules (WFRs). For example amo (to love) and amator (lover) are connected with a relationship that describes a change from a verb to a noun through the addition of a suffix (-a-tor) that in itself bears semantic information (in this case it characterises agentive and instrumental nouns, i.e. someone or something performing an action).

WFL has received funding from the European Union's Horizon 2020 research and innovation programme (Marie Sklodowska-Curie grant agreement No 658332-WFL). The resource is still a work-in-progress - having so far covered 5,366 morphological families, 268 WFRs, 22,679 relations - and is due to be completed by October 2017. The lexical basis used for the resource comprises the whole 69,682 lemmas featured in the morphological analyser for Latin LEMLAT 3.0.

The word formation lexicon is built in two steps:

1. Word formation rules (WFRs) are detected using a mixture of previous literature on Latin derivational morphology (Jenks, 1911; Fruyt, 2011; Oniga, 1988) and semi-automatic procedures (Passarotti and Mambrini 2012).
2. WFRs are applied to lexical data: lemmas and WFRs are paired using a MySQL relational database, and a number of MySQL queries provide the candidate lemmas for each WFR. Input and output pairs are then checked manually, in order to clear out false friends and duplicate results due to homography.

This poster will describe the resource and illustrate the web application that is being developed to easily access the data.

The WFL dataset is both integral part of Lemlat and used in a standalone web application. The database will be made available for download, so that extensive queries can be run and the data can be used and reused at will. The web application is intuitive and user-friendly. It supports those scholars and students that are not familiar with database querying languages such as SQL, but also Classicists with specific scientific questions.

The lexicon can be browsed either by WFR, affix, input and output Part-of-Speech (PoS) or lemma. Drop-down menus provide the available options for each selection, such as the list of affixes and lemmas. Results are visualised as lists of lemmas and tree graphs, whose nodes are lemmas and edges are WFRs. Trees are interactive. Clicking on a node shows the full derivational tree ("word formation cluster") for the lemma reported in that node. For example, figure 1 shows the word formation cluster for the lemma computo, 'to calculate'. Clicking on an edge shows the lemmas built by the WFR described by that edge. Methodological motivations will be given for each browsing option together with suggestions for potential uses of the web to investigate Latin derivational processes. Four browsing choices can help the scholar with an array of linguistic investigations.

1. By WFR - opens research questions on a specific word formation behaviour; for example, it is possible to view and download a list of all verbs that derive from a noun with a conversive derivation process (e.g. radix 'root' -> radicor 'to grow roots').
2. By Affix - acts similarly as above, but works more specifically on affixal behaviour: for example, it is possible to see all agentive nouns in -tor and verify how many correspond to a female equivalent in -trix.
3. By PoS - useful for studies on macro-categories, such as nominalisation or verbalisation.
4. By Lemma - useful when studying the productivity of one specific morphological family (like the one for bellum above) or a group of morphological families.

These explorations lead in many directions through investigations on derivational production and semantics (Can semantic identification of outputs help to show which WFRs are more morphotactically transparent? Which inputs produce a certain kind of outputs? Etc.).

The poster will illustrate a few applications of the resource and a demonstration of case studies. The poster will be accompanied by a live demonstration.

Bibliography

Forcellini, A. (1940) "Lexicon totius latinitatis ab Aegidio Forcellini seminarii Patavini alumno lucubratum, deinde a Iosepho Furlanetto eiusdem seminarii alumno emendatum et auctum, nunc vero curantibus Francisco Corradini et Iosepho Perin seminarii Patavini item alumnis emendatius et auctius melioremque in formam redactum." Tom. I AG, Patavii, Typis Seminarii

Georges, K.E. (1880) Ausführliches lateinisch-deutsches und deutsch-lateinisches Handwörterbuch... Vol. 2. Hahn'sche Verlags-buchhandlung, 1880.

Glare, P.G.W. (1982) Oxford latin dictionary. Clarendon Press. Oxford University Press, 1982.

Gradenwitz, O. (1904) Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas. S. Hirzel.

Jenks, P. R. (1911) A manual of Latin word formation for secondary schools. DC Heath & Company, 1911.

Fruyt, M. (2011) "Word-Formation in Classical Latin." A Companion to the Latin Language. 157-175.

Oniga, R. (1988) I composti nominali latini: una morfologia generativa. Vol. 29. Pàtron.

Passarotti, M.C. (2004). "Development and perspectives of the Latin morphological analyser LEMLAT." Linguistica computazionale 20, no. A. 397-414.

Passarotti, M, and Mambrini, F. (2012) "First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin." In Eighth International Conference on Language Resources and Evaluation, (LREC 2012), pp. 852-859. European Language Resources Association (ELRA).

American Panorama: An Atlas of United States History

Justin Madron
jmadron@richmond.edu
University of Richmond, United States of America

Nathaniel Ayers
nayers@richmond.edu
University of Richmond, United States of America

Introduction

The poster will focus on the Panorama Toolkit built by the Digital Scholarship Lab at the University of Richmond (DSLUR) and Stamen Design. The first section will outline the capabilities of the toolkit then turn to examples from DSLUR's American Panorama project.

The focus of the toolkit is on the richly dynamic capabilities of modern web maps to enable deep digital scholarship on American History. The components are built on the foundation of open source tools including React and Leaflet libraries -- creating an extensible, public, collaborative framework that will support the continued development of future maps in our Atlas or the user's maps of choice. The components within can be installed via npm and integrated into any web-facing project. American Panorama also includes a template that can be used as a starting point for maps in the American Panorama Atlas, or for other projects that aim to use Panorama components. All of the components are "views" (meaning they appear in the DOM) and use React.

Since it is open-source, people can contribute and add new components that can be used individually, or wired together to create dynamic visualizations. The flexibility of the toolkit allows for a range of complexity, which has been seen in the five maps currently available. We envision this being used not only in academia, but in other projects varying in subject and purpose.

To highlight the possibilities of the toolkit, we will focus on University of Richmond's American Panorama, a series of online maps that use interactive mapping features and innovative cartographic design to encourage active investigation of American History. For over a year, the DSLUR has been working with Stamen Design to investigate spatial exploration across various subjects in 19th and 20th century U.S. history. These maps—which make up the Atlas—cover a wide range of topics designed to communicate knowledge about some facet of American history to audiences that include students, scholars, and the general public.

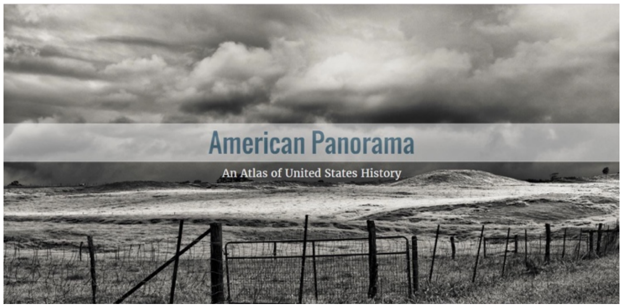
In 2015, Stamen and the DSLUR collaborated on the first four maps of the Atlas. The first map covers the forced migration of enslaved people before the Civil War, highlighting the change in net importation and exportation of slaves to and from different parts of the South for each decade between 1810-1860. The migration across the Overland Trails to the West focuses on the routes emigrants took envisioned through the writings in their journals, along with the locations of forts, ferries, and bridges. The movement of people and goods through canals highlights not only the location of canals, but visualizes the goods that moved along them. The immigration of people to the U.S. from 1850 to today shows Foreign-Born populations at various scales from country level to county level.

The collaboration between Stamen and DSLUR was the key driver to the success of the first four maps. We combined the knowledge and narrative ideas of DSLUR and the visualization skills of Stamen to translate new knowledge into inviting, elegant displays that respect the complexity of the subjects. American Panorama combined two deeply passionate teams to create not just four deeply engaging maps, but also a toolkit of reusable visualization components that could be used for future atlases in the American Panorama.

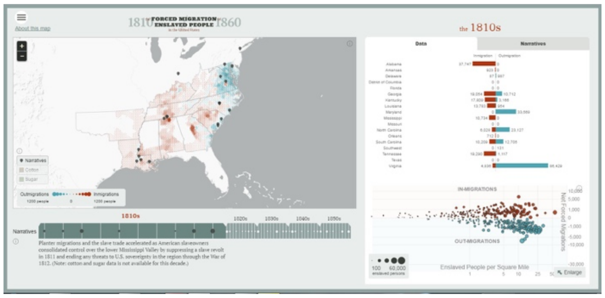
In 2016, DSLUR employed the Panorama toolkit to build the fifth map in the American Panorama series, a visualization of the Home Owners' Loan Corporation (HOLC) redlining maps from the early 20th Century. Using the visualization tools developed by Stamen, DSLUR created Mapping Inequality, which updates the study of New Deal America, the federal government, housing, and inequality with twenty-first century tools and scholarship. It offers unprecedented online access to the United States government's collection of "security maps" and area descriptions produced between 1935 and 1940. The toolkit allowed the use of cartographic visualizations like concentric circles, which arguably were the spatial episteme of cities for early twentieth-century Americans. These Concentric circle diagrams visualize the relative distribution of HOLC grades in relation to the center of the city.

The toolkit's capabilities are seen in the fifth Atlas map, Mapping Inequality. The foundational components, as well as the flexibility of the software, allowed DSLUR to create an engaging, invoking map highlighting the practice of denying access to mortgages and other forms of credit to neighborhoods based upon discriminatory practices, which later became known as redlining. This last installment of the Atlas showcased that this can be used by anyone wanting to create a project using the capabilities of modern web mapping seen in American Panorama.

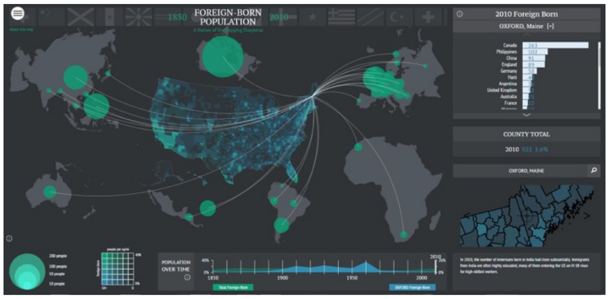
Screenshots



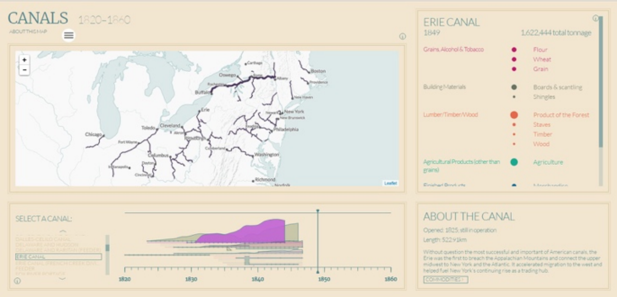
American Panorama



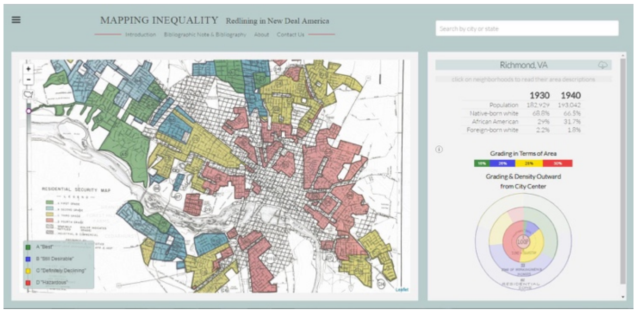
Foreign-Born Population



Forced Migration of Enslaved People



Canals



Mapping Inequality: Redlining in New Deal America

Bibliography

Digital Scholarship Lab, University of Richmond(2015–) American Panorama. *Dsl.richmond.edu/panorama

American Panorama. (n.d.) Panorama - Github repository: https://github.com/americanpanorama/panorama

Stamen Design LLC. (n.d.) Stamen. *stamen.com

Between Page and Screen by Amaranth Borsuk and Brad Bouse: Translating an AR poetry book

Piotr Marecki
piotr.marecki@ha.art.pl
Jagiellonian University, Poland

Aleksandra Małecka
aleksandra.malecka@ha.art.pl
Jagiellonian University, Poland

Our poster is devoted to the English-Polish translation of the augmented reality book *Between Page and Screen* by Amaranth Borsuk and Brad Bouse. The prominent digital media scholar K. N. Hayles wrote about the relationship between the analogue and the digital: “In the contemporary era, both print and electronic texts are deeply interpenetrated by code. Digital technologies are now so thoroughly integrated with commercial printing processes that print is more properly considered a particular output form of electronic text than an entirely separate medium.” *Between Page and Screen* is an illustration of these processes on one of the possible reading levels. The authors tell the tale of the relationship between the analogue book and digital screen in the form of an epistolary concrete poetry dialogue between two protagonists, P. and S. The nature of the relationship is underlined by the fact that the work comprises of a traditional book with QR codes (sold in bookstores) and an online application for reading it. The two platforms are inextricably joined in the reading process (which cannot take place if one of the links is missing). The translation of the book and application involves innovations and challenges unheard of in the case of conventional books. Our poster is a visual guide through the process of translation and publication of the AR book in Poland.

Creating the data base

Borsuk and Bouse’s work can be seen as a piece of constrained writing. The authors tell their story with a selected vocabulary. The letters by P. and S. highly employ words that stem from the Indoeuropean roots of the words “page” and “screen”. The authors used lists of related words found in the respective entries in the *American Heritage Dictionary of Indo-European Roots*. In the Polish translation, the translators recreated the same principle for the Polish words “kartka” and “ekran”, but given that there is no one comprehensive dictionary of Indo-European roots for the Polish language, they created similar lists based on the study of several etymological dictionaries of Slavic languages. The etymological and lexicological research was the first, crucial step of the translation process.

Comparing the data bases

Given different linguistic contexts and different lists of related words, it becomes clear that the dialogues and characters of the analogue and digital protagonist (in Polish K. and E.) will differ between the source and target language. It was thus necessary to reinvent the characters based on the new vocabulary ranges.

Conceptual translation

Given that word-for-word translation was impossible with the above-mentioned constraint, the translators decided to recreate the logic of the book, loosely following the plot of the English original. The final translation is rather a translation of the concept than a direct translation of the text.

Visual adaptation

The work by Borsuk and Bouse can be also seen as concrete poetry. The QR codes in the book have a visual aspect to them themselves. In interaction with the online application, they launch animations, which in their shape and movement reflect the verbs or nouns used in them. Given that the word resources of the Polish and English works are different, the shapes of the animations were also changed.

Use of software

The application was written with the FLARToolkit, thus, any changes, from content to features like diacritics, were introduced with the use of this tool.

Publication and distribution

The last element of this procedure is traditional computer typesetting of the book and introducing it into distribution as well as launching a website dedicated to it.

The translation of electronic literature remains an understudied area. One of the reasons for this is the acceptance of the status quo that most available, created and read works are in English. This is connected with several difficulties in translation and adaptation of electronic literature (working with code, working with a given platform). There are often several issues connected with the idiomatic character of the work (both on the level of content and platform). Our poster also does not offer a ready formula for the translation of such works – it is a case study of one work, considered emblematic of the phenomenon of AR books. The methodologies of medium specific analysis (K. N. Hayles), as well as exploratory programming (N. Montfort), the terms bookishness (J. Pressman) and liberature (Z. Fajfer and K. Bazarnik) provide the vocabulary that will allow us to describe and visualize the performed process.

Bibliography

Bolter J. D., (2014), *Augmented Reality*. In: *The John Hopkins Guide to Digital Media*, M.-L. Tyan, L. Emerson, B. J. Robertson (eds.), John Hopkins University Press: Baltimore.

The Creative Computing Lab in Poland

Piotr Marecki
piotr.marecki@ha.art.pl
Jagiellonian University, Poland

Jakub Woynarowski
j.woynarowski@gmail.com
Academy of Fine Arts Krakow, Poland

The poster is devoted to the project we are currently working on at the Jagiellonian University with digital media artists, programmers, and theorists: the creative computing laboratory. The main aim of this team is to research the creative process in its widest definition in the era of digital textuality. Digital literature is the focus of our study, above all highly computational literature. In the laboratory environment a team of artists, theorists, and programmers are working on the forms of contemporary creative programming, closely tied to the choice of programming language and the platform on which the work is written. One of the project's aims is to create digital artwork in the laboratory system, and to produce an academic description of the process. The invited artists, theorists, and programmers are gathered in three spheres of creative programming: text generators, interactive fiction, and experiments with augmented reality. One of the outputs of the project is a catalog and an archive of programed works, both for research and for teaching. The team is also focusing on the phenomenon of homegrown “writing under constraint” techniques, which are seen as precursors to creative programing techniques. As part of our search for original and undiscovered phenomena in the humanities, we are researching the local demoscene as an example of creative teamwork in digital media. This project is being carried out as a transdisciplinary task, at the crossroads of computer studies, literary studies, creative writing, experimental literature, and digital media. It also develops new tools in humanities research (laboratory teamwork) and new genres and communication practices (technical reports, open notebook science). As part of the project we are producing expressive processing (presupposing knowledge of the code) and platform studies (presupposing thorough knowledge of the capabilities of the platform where the works are written) tools for describing digital works. One of the aims of the project is to create a retro platform archive, as well as an archive of Polish works written for various platforms.

The poster will present the team's one year experience (the lab was launched in May 2016). We will focus on the most important terms in the dictionary useful for the lab's team, methods developed in collaboration between theorists, artist and programmers working together in the academic context. Moreover, we will build a guide on how to build such an institution in the geopolitical, economic and cultural context, which determine the scope of our research, approach, structure and significance of the lab. By describing our lab, we try to tackle trends that are relevant to contemporary studies on digital media, taking into account and affirming the local perspective, which is different from the dominant one.

Bibliography

Montfort, N., (2016), *Exploratory programming for the Arts and Humanities,* Cambridge: MIT Press.

Wardrip-Fruin, N. (2009), Expressive Processing. *Digital Fictions, Computer Games, and Software Studies.* Cambridge: MIT Press.

The Victoria Press Circle

Miranda Marraccini
mcm5@princeton.edu
Princeton University, United States of America

The Victoria Press began as an outgrowth of SPEW (Society for Promoting the Employment of Women), a group of Victorian feminists who sought to provide new avenues for women's work in the printing industry. SPEW activists, led by Emily Faithfull, set up a Press where women worked as compositors (Tusan, 2004; Fredeman, 1974). Presswomen printed anthologies, tracts, and feminist periodicals, including the monthly English Woman's Journal (1858-1864), which published mainly contributions by women. Publication of the English Woman's Journal (EWJ) spanned "the period between the failed attempt to reform legislation that prevented married women holding property in 1857 and the equally unsuccessful attempt to win female suffrage in 1867" (Mussell, 2008). Articles in the EWJ promoted both of these reform measures and advocated for female employment in different fields, as well as other contemporary feminist causes. My use of the term "feminist," though anachronistic to the movement, has been accepted by other scholars writing about the period, who choose the word to demonstrate an active and purposeful engagement with the continuous struggle for women's rights (see Phegley, 2004; Mussell, 2012; Frawley, 1999).

My digital project, The Victoria Press Circle, funded by Princeton's Center for Digital Humanities, offers open-access network visualizations of the women and men involved in the Victoria Press, based on contents of the EWJ and three anthologies printed at the Press between 1861 and 1863. The Victoria Press Circle's first aim is reconstruction: the project helps to establish the history of the Victoria Press, since there is no existing archive. This is especially important since the EWJ includes a high percentage of unsigned contributions (about 40 percent). None of the women who published in the EWJ currently have significant digital representation. Identifying them as individuals combats the critical undervaluing of texts in female-produced periodicals, and studying them as a group highlights authors who may not receive attention individually.

Furthermore, this project demonstrates collaboration. A network-focused approach is particularly appropriate because the Victoria Press was constructed on a material model of collaborative female labor. Its founders explicitly attempted to build a hub of social interaction around the Press, creating venues to promote women's rights. SPEW members saw their office as a meeting place for women advocating for female employment:

It is also the intention of the Society to render their office a depôt for information of every kind relating to the employment of women. Curious and interesting facts will be collected. Extracts from newspapers, pamphlets, and speeches on the subject, will be gathered together, and kept for the inspection of members of the Society (EWJ, 1859).

In creating network graphs, I am reconstructing how the Presswomen built a social network for themselves, not imposing intentionality on their project (Weingart, 2013). All of my data will be freely available and downloadable in .csv format for other researchers to access and use.

Ultimately, The Victoria Press Circle's open-access website will display at least three network visualizations, constructed in Cytoscape, of those involved in the Press: one composite graph for the three anthologies; one graph for the EWJ; and one combined visualization for all the publications. Cytoscape, though a tool designed for biomolecular analysis, is more flexible than Gephi for social network analysis, especially for specific functions of filtering based on node and edge attributes and on network statistics (see Shannon, 2003).

In addition to literary contributors, my visualizations include compositors, engravers, printers, editors, and paper manufacturers. Marianne Van Remoortel (2015) has helped in identifying names of compositors from newspaper reports and census data. While the individual model of many digital archives privileges authors and minimizes others involved in literary production, the women of the Press were working at every level of print culture to advance their social aims. By valuing all types of contributions equally, my visualizations illustrate their collaborative effort. The Presswomen's project echoes through current debates in digital humanities about the necessity of learning to code for engagement in DH work (Dinsman, 2016). I believe that programming can be a feminist act for scholars, just as involvement in print culture was a feminist act for Victorian female authors. The artisan practice of printing is analogous to the artisan practice of coding, and both are affected by experiences of gender, race, class, and sexuality. I hope to use my project to show how women worked with their hands and their pens in tandem, and how I'm continuing that work today.

Bibliography

(---) (1859).“Association for Promoting the Employment of Women,” English Woman's Journal IV(19): 59. http://ncse-viewpoint.cch.kcl.ac.uk/ (accessed 10 September 2016).

Dinsman, M. (2016). “The Digital in the Humanities: An Interview with Marisa Parham.” Los Angeles Review of Books. https://lareviewofbooks.org/article/digital-humanities-interview-marisa-parham/ (accessed 8 February 2017).

Frawley, M. (1999). “Feminism, Format, and Emily Faithfull's Victoria Press Publications.” Nineteenth-century feminisms 1: 39–63.

Fredeman, W. (1974). “Emily Faithfull and the Victoria Press: An Experiment in Sociological Bibliography,” The Library s5–XXIX(2): 1439-164. doi:10.1093/library/s5-XXIX.2.139 (accessed 13 March 2016).

Mussell, J. (2008). “NCSE: English Woman's Journal (1858-1864).” http://www.ncse.ac.uk/headnotes/ewj.html (accessed 10 September 2016).

Mussell, J. (2012). The Nineteenth-Century Press in the Digital Age. Palgrave Studies in the History of the Media. New York: Palgrave Macmillan.

Phegley, J. (2004). Educating the Proper Woman Reader: Victorian Family Literary Magazines and the Cultural Health of the Nation. Columbus: Ohio State University Press.

Van Remoortel, M. (2015). Women, Work and the Victorian Periodical: Living by the Press. New York: Palgrave Macmillan.

Shannon, P. et al. (2003). “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks,” Genome Research 13(11): 2498–2504. doi:10.1101/gr.1239303 (accessed 15 October 2016).

Tusan, M. (2004). “Performing Work: Gender, Class, and the Printing Trade in Victorian Britain,” Journal of Women's History 16(1): 103-126. doi:10.1353/jowh.2004.0037 (accessed 10 September 2016).

Weingart, S. (2013). “Networks Demystified 8: When Networks Are Inappropriate,” The Scottbot Irregular. https://scottbot.net/networks-demystified-8-when-networks-are-inappropriate/ (accessed 15 October 2016).

Humanities At Scale – Creating A Network Of Expertise And A Sustainable Infrastructure For Digital Humanities Projects In Europe

Markus Matoni
matoni@sub.uni-goettingen.de
State and University Library Göttingen, Germany

Jenny Oltersdorf
oltersdorf.jen@gmail.com
FH Potsdam, Germany

Dirk Roorda
dirk.roorda@dans.knaw.nl
Data Archiving and Networked Services, Netherlands

Digital Humanities (DH), as a growing field, generates lots of innovative new tools and methods to enhance work in humanities disciplines. To ensure a sustainable digital European infrastructure for long-term availability of digital research data and tools the DARIAH infrastructure (Digital Research Infrastructure for the Arts and Humanities) aims at the development of a strong cooperation and information exchange between research communities and institutions in the arts and humanities. The ambition to extend this community and create a European-wide network of expertise was one of the reasons to initiate the new Horizon 2020 project Humanities at Scale (HaS).

The project coordinated by DARIAH-EU aims to evolve the DARIAH community, offer training and information material, set up a number of training workshops and develop core services within a sustainable framework. Within this framework a research infrastructure to connect DH research tools, services and data will be established and the challenge is to build technical systems to support this.

The poster will highlight and answer the following questions from a technical perspective:

- How can a decentralized framework (architecture of participation) of tools and services be developed and maintained?
- Which kind of technologies are useful to establish a sustainable framework for DH-tools and -services?
- What is needed to integrate externally developed tools into such a framework?

The DH community, as a network of expertise, created and still creates a lot of important tools and services - often only in a specific project context or for one particular research question. To make these tools and services accessible, re-usable, and open for DH, HaS is building a sustainable technical framework which will advance the creation and enhancement of digital tools. This will improve applications, enlarge possibilities for publication and enable long-term preservation of research data and digital methods. Just hosting digital tools is not sufficient. Code and server configuration require constant maintenance and ongoing development - an elementary issue for a decentralized and unified infrastructure. Thus, within the HaS project, researchers and developers get conceptual and technical support for their tools and services to enable a better integration into the community.

Furthermore, projects in the Digital Humanities need key central services. They often must spend a large proportion of their available resources on basic infrastructure components. To investigate demands regarding service requirements and tools a survey was conducted in summer 2016. Its aim was to get an overview of the currently used services and tools and the ones, which will become relevant in the future. The survey design enabled a gap analysis focusing on the following three kinds of tools and services:

- Development services such as Git/SVN, Chili or Open Project. They support developers in code hosting, code review, the development of workflows and documentation. They improve the organization and quality of code and the workflows for writing it.
- Hosting services comprise virtual machines, storage and repository solutions, configuration management tools and monitoring environments. They provide basics for a sustainable development and ongoing operation of DH services.
- Collaboration tools support the whole area of DH. Collaborative writing and managing of digital content in a content management system are major ingredients of everyday's work.

To connect all services and tools HaS will define requirements to enhance the integration.

Besides these enabling, generic tools, the DARIAH community is also producing highly specific research tools that we need to take into account. That will be organized as a bottom-up process, drawing information from existing tools that have been contributed to DARIAH or tools which are offered as candidate contributions. In order to harvest their characteristics, HaS is building a contribution registry, called the DARIAH Inkind Tool. It should not only collect technical specifications, but also generate an overview of what is available in terms of research capabilities. And last but not least, the registry will be used to assess the cost and value of the DARIAH contributions. The existence of the Inkind registry will enable the DH community to muster a lot of information about each tool, such as documentation, use-cases, application descriptions, data examples, research scenarios and particular training possibilities.

The general aim of the poster is to present the results of the project to the community to get important feedback. This includes the chance to present the participants of the poster-sessions a live-demo (German and English) of the contribution registry. All in all, the presentation will provide insights on how we face sustainability from a technological view.

Designing from a Narrative: Leveraging the 6 Point Story Method to Facilitate Interaction Design in Digital Humanities Projects

Shane McGarry
shane.mcgarry.2015@mumail.ie
Maynooth University, Ireland

One of the many challenges facing Digital Humanities projects today is how to design for alternative approaches to traditional close reading. Many of the current interface mechanisms for representing text are drawn from their print counterparts and are directly reflective of the analogue. These mechanisms rely upon traditional representations of text and typically structure content using techniques that have been leveraged by print for several centuries. However, these various mechanisms are designed to support traditional close reading techniques and often fail to consider digital reading methods. As digital content has become more ubiquitous and commonplace, a demographic shift has occurred, migrating readers towards alternative methods of reading and engagement with text, alternatives which seek to augment the traditional close reading approach. Previously conducted research shows a decline in traditional close reading techniques when moving from print to digital (Hayles, 2012; Mangen, 2008). This, along with other factors, has led to the emergence of new methods of reading and textual engagement, such as hyper reading, radial reading, and distant reading (Hayles, 2012; McGann, 2012; Moretti, 2013). However, finding metaphors or interactions which support these different modalities can be challenging, and the methods for discovering how users leverage content can often be inconsistent.

In an effort to increase user experience within the Digital Humanities space, the field has turned to interaction design techniques in order to better understand how the user leverages these types of projects (Ruecker et al, 2011; Pierazzo, 2014). Traditionally, interaction design has borrowed from social science in an effort to create focus groups or usability studies which provide insight into how the user interacts with the software in question. Various techniques—such as ethnography, surveys, and interviews—are often leveraged with the goal of understanding how the individual engages with the software. However, these techniques typically fail to consider why the user may engage with the software in question. This assessment, otherwise known as the “emotional impact” of the software, can have tremendous impact upon the reader’s interaction with the content. Studies have shown that the subjective experience often has a strong impact upon metacognition (Ackerman & Goldsmith, 2011), leading readers to demand more than just efficiency and usability from a platform: they also require emotional engagement (Shih and Liu, 2008). This poster proposes the “6 Point Story Method” as a framework which can be used to assess not only how the reader seeks to interact with the content but also why.

The “6 Point Story Method” is a psychological method developed by Alida Gersie. The method is primarily used by psychologists as a therapy modality to assist patients by framing stories which often have subconscious meaning and acting as an assistant to uncover issues requiring further therapeutic intervention (Dent-Brown and Wang, 2004). This poster proposes that this method can be adopted for use in the early stages of interaction design with the goal of facilitating idea generation in a group setting in order to further understand: the goals of end users, what challenges the users may face when navigating a software system, and the driving motivation behind the use of the platform (thus forming the emotional component of the analysis). This poster will discuss the use of the 6 Point Story Method within a focus group setting whose primary goal is to further explore how users wish to engage with Digital Humanities projects from non-traditional engagement methods (e.g. in support of reading modalities outside of traditional close reading). The methodology, including the demographics behind the focus groups, will be illustrated and the results will be explored in order to highlight the advantages of adapting the 6 Point Story Method to support traditional interaction design techniques.

Bibliography

Ackerman, R. & Goldsmith, M., (2011). “Metacognitive Regulation of Text Learning: On Screen Versus on Paper”. Journal of Experimental Psychology, 17(1): 18-32.

Dent-Brown, K. & Wang, M., (2004). “Developing a rating scale for projected stories”. Psychology and Psychotherapy, 77(Pt 3): 325-333.

Hayles, K.N., (2012). How We Think: Digital Media and Contemporary Technogensis, Chicago: University of Chicago Press.

Mangen, A., (2008). “Hypertext Fiction Reading: Haptics and Immersion”. Journal of Research in Reading 31(4): 404-419.

McGann, J., (2001). Radiant Textuality: Literature After the World Wide Web, New York, NY: Palgrave.

Moretti, F., (2013). Distant Reading, London: Verso.

Pierazzo, E., (2014). Digital Scholarly Editing: Theories, Models, and Methods. New York, NY: Routledge.

Ruecker, S., Radzikowska, M., & Sinclair, S., (2011). Visual Interface Design for Digital Cultural Heritage. Burlington, VT: Ashgate

Shih, Y. & Liu, M., (2007). “The Importance of Emotional Usability”. Journal of Educational Technology Systems 36(2): 203-218.

Mapping Violence: Visualizing State-Sanctioned Racial Violence on the Mexico/Texas Border (1900-1930)

Jim McGrath
james_mcgrath@brown.edu
Brown University, United States of America

Monica Martinez
monica_martinez@brown.edu
Brown University, United States of America

Cole Hansen
cole_hansen@brown.edu
Brown University, United States of America

Edward Jiao
edward_jiao@brown.edu
Brown University, United States of America

This poster will describe recent work on *Mapping Violence,* a digital project that makes visible decades of state-sanctioned acts of violence in the United States during the first decades of the twentieth century. From 1900 through 1930, vigilantes and Texas Rangers killed hundreds of ethnic and national Mexicans along the Mexico/Texas border. "A man's life just wasn't worth much at all," recalled eyewitness Roland A. Warnock, "There were so many innocent people killed in that mess that it just made you sick to your heart to see it happening." This period of conflict is one of the largest episodes of civil unrest in American history and still impacts the struggle for justice and civil rights that continues today in Texas and in America at-large. However, this history is widely unknown. Like *Visualizing Emancipation* (University of Richmond), *Digital Harlem* (University of Sydney), and other projects that remind us of the hyperlocal stories and histories of particular geographic regions, *Mapping Violence* aims to recover and make accessible lives, events, and acts of oppression that continue to impact later generations in various ways.

Mapping Violence is part of *Refusing to Forget*, a public humanities initiative involving collaborators from Brown University, Loyola, Texas A&M, the University of Texas, and

South Texas College (among others) that is working to raise awareness about this history of racial violence and its legacies. *Mapping Violence* is complemented and informed by additional endeavors by the *Refusing to Forget* team: academic articles and books, applications for Texas historical markers acknowledging this history, collaborations with Texas public school educators to develop curriculum, and exhibitions of material with the Bullock Texas State History Museum. In *Mapping Violence*, we envision a project that functions as a database of primary and secondary sources for researchers studying this history, a curated and dynamic resource for classroom use, a memorial space that honors and remembers the victims of these acts of violence, and a corrective to historical narratives and commemorations that erase or mute these moments in the history of Texas, Mexico, and the United States.

This poster will document recent and ongoing work at Brown University on the completion of the first public-facing iteration of *Mapping Violence* (currently scheduled to launch in the fall of 2017). It will provide an overview of the creation of a database of material derived from primary and secondary texts, narrate the iterative processes of mapping and annotating data about state-sanctioned violence, and consider the ways in which the design of the project is informed by its stakeholders and its range of imagined audiences (educators, academics, activists, students, among others). The poster will highlight the significant roles undergraduate collaborators (from a range of academic disciplines including Computer Science and Ethnic Studies) have played in the project at various stages: as researchers, programmers, designers, developers, and curators. We will also document the project's timeline, workflow, and challenges. What are the challenges inherent in creating a digital space that is both an act of commemoration and a source of research (challenges faced by projects with similar inclinations like *Our Marathon: The Boston Bombing Digital Archive* and *The 9/11 Digital Archive*, among others)? In what ways can database creation and digital curation productively remediate and recontextualize controversial historical events for public audiences? How do we navigate desires to tell hyper-specific, localized stories about individual victims of violence with inclinations to inventory and resituate these events within larger narratives about race, gender, and nationalism?

A working demo of *Mapping Violence* will accompany the poster, provided there is room and available resources to do so.

Public Humanities: In Search of a Field

Jim McGrath
james_mcgrath@brown.edu
Brown University, United States of America

Robyn Schroeder
robyn_schroeder@brown.edu
Brown University, United States of America

Inge Zwart
inge_zwart@brown.edu
Brown University, United States of America

Like the phrase "digital humanities," "public humanities" has proven difficult to pin down, even by practitioners who have come to value this term as an accurate description of their research interests and institutional affiliations. It has been a term used (primarily, but not exclusively, by institutions of higher education in North America and by grant-funding agencies like the National Endowment for the Humanities) to describe various forms of cultural production: physical and digital acts of curation, archival initiatives, community outreach, educational programming, and public arts initiatives. It has ties to terms like "public history." More recently, investments in "public humanities" have more frequently intersected (and at times elided) investments in "digital humanities," given the public-facing imperatives of many digital projects and practitioners. Are all digital humanities projects also inevitably public humanities initiatives? When the phrase "public humanities" is invoked, who is most often speaking, and to what do they refer? Who is most likely to fund projects with explicit ties to this term, and what communities are served and financed by these efforts? When practitioners of public humanities write about the value of their work, who do they cite? More generally, what forms, contexts, people, places, geographic regions, occupations, and communities is the term "public humanities" tied to, and who is missing, obscured, erased, or otherwise apart from these conversations?

This poster will document recent efforts by public humanities practitioners at the John Nicholas Brown Center for Public Humanities and Cultural Heritage (Brown University) to survey and reflect upon both the ten-year history of the Center (and its M.A. program in Public Humanities) and, more generally, the state of "public humanities" as a field of critical inquiry. Postdoctoral researchers and graduate students at the Center have determined that the institutional memory of the Center would benefit greatly from an indexed archive of primary materials (syllabi, photographs, exhibition materials, student projects and writings, among others) as well as a collection of oral histories from major figures in its history (directors, staff, faculty, graduates, community fellows and collaborators), and we are in the process of building this collection in collaboration with representatives of Brown's Center for Digital Scholarship and the Brown Library's digital repository service. We have also begun assembling a database of texts (books, journal articles, grant application guidelines, blog posts, and the language of Center / program recruitment literature, among others) to better understand where and why the phrase "public humanities" has been utilized between the late twentieth century (starting around 1970, in Congressional documents that informed the creation of the NEH) and the early twenty-first century. The poster will contextualize the motivations behind our decisions to focus on particular corpora and the organization of our database. Beyond the first iteration of this hyperlocal history (which will materialize in the form of the first stage of the Center's collection in Brown's digital repository and will be publicly available to researchers), we will concurrently release a dataset of public humanities corpora and a series of visualizations exploring that data. One will begin to map the geographic regions and institutional ties to public humanities programs and grant-funded public humanities initiatives. Another will document part of the citational history of public humanities discourse by visualizing links between texts present in bibliographic data of a sample set of academic texts. Representative samples of these visualizations, which are not meant to be comprehensive so much as they are intended to be models for potential avenues of inquiry, as well as the methodologies informing them, will be featured in the poster. We will also link these datasets to the crowdsourced data we receive from "Day of Public Humanities" (#DayofPH), a digital project held in May of 2017 that, like the "Day of Digital Humanities" initiative, aims to make the various forms of day-to-day labor (and laborers) in the field visible and to encourage dialogue across and beyond these contexts.

ARL Digital Scholarship Institute

Sarah Melton
sarah.melton@bc.edu
Boston College, United States of America

Michelle Dalmau
mdalmau@indiana.edu
Indiana University, United States of America

Nora Dimmock
ndimmock@library.rochester.edu
University of Rochester, United States of America

Dan Tracy
dtracy@illinois.edu
University of Illinois at Urbana-Champaign
United States of America

Erin Glass
erglass@ucsd.edu
UC San Diego, United States of America

This poster will reflect on the Association of Research Libraries' (ARL) upcoming inaugural week-long Digital Scholarship Institute for library professionals. Scheduled for June 2017 at Boston College, the Institute will introduce librarians and staff who are not currently involved in digital scholarship to the methodologies and considerations of such work. This poster will detail successes and lessons learned for future Institutes.

Following two large-scale workshops hosted by the Coalition for Networked Information (CNI) on planning and supporting digital scholarship centers in 2014 and 2016, staffing and training emerged as fundamental components to successful initiatives. In 2014, workshop participants agreed that:

> *"...it is more important to have the proper mix of abilities overall across personnel than it is to have any one particular type of staff member. The ability to learn new skills, adaptability, and agility are qualities in personnel that can be even more important than the expertise that they initially bring to the position."* (Lippincott and Goldenberg-Hart, 2014:7)

In 2016, CNI hosted a follow-up workshop in partnership with ARL with over 100 participants. The resulting report from that meeting made it clear that *people* are core to the success of digital scholarship initiatives. Several points of consensus emerged from the 2016 workshop, two of which ARL has adopted for the Digital Scholarship Institute:

- Staff need time including time for research, space, recognition, and funding for intensive professional development, and the freedom and ability to develop library digital projects
- Skill-building goes beyond workshops and collaborations, and is often rooted in culture change and support from administration (Goldenberg-Hart, 2016.)

In response to these discussions, the ARL Digital Scholarship Institute Advisory Group identified several core principles that underpin digital scholarship initiatives with an emphasis on cultural change: collaboration, critical thinking, and disciplinary contexts. Rather than starting with tools or technologies, the Institute emphasizes these concepts as building blocks of digital scholarship.

By providing a five day, face-to-face immersive environment, the Institute will cultivate collaboration and create cohorts of digital scholarship practitioners. These cohorts will learn together, share experiences, and work on future endeavors through follow-up online meetings that will ensure participants have opportunities to put core principles into practice. Participants will learn core principles related to metadata and data curation, copyright and fair use, and project management. They will also learn how to apply these concepts in the context of digital pedagogy and collaboration with researchers. These applications build on existing librarian skillsets related to instruction and to work with researchers needing to clarify and refine research ideas. The framework of the Digital Scholarship Institute thus bridges existing library capacities with new knowledge and skills that will transform their work as they increasingly partner with researchers on digital scholarship using diverse digital methods and tools.

The ARL Digital Scholarship Institute also builds on the University of Rochester's pilot Institute for Mid-Career Librarians in Digital Humanities funded by the Andrew W. Mellon Foundation. The program invited twenty librarians from across the US and Canada to take part in a three-day intensive residential institute followed by monthly online meet-ups from July 2015-June 2016.

The pilot successfully built a strong cohort among the attendees, who were assigned to one of four curriculum tracks based on their interests: text encoding, analysis and visualization; digital pedagogy and digital media literacy; digital mapping and archeology; and metadata, data curation, and data modeling. A core curriculum included workshops on project management, data modeling, and copyright/fair use. Objectives of the pilot institute included strengthening librarian practice to support digital humanities, creating a structure to support sustained skills development, and integrating with future ARL initiatives.

Attendees participated in assessment activities throughout the year-long pilot, providing data to inform the creation of sustainable programs. In October 2016 teams from five institutions, University of Rochester, Indiana University, University of Illinois at Urbana-Champaign, San Diego State University, and Boston College, met to workshop the creation of a program that would be foundational to creating these new communities of practice. This multi-institutional initiative will provide an opportunity for broader transformation in the academic library ecosystem at the level necessary to create a strong community of practice around digital scholarship work.

Bibliography

Goldenberg-Hart, D. (2016). "Planning a Digital Scholarship Center 2016." Coalition for Networked Information, https://www.cni.org/wp-content/uploads/2016/08/report-DSCW16.pdf.

Lipponcott, J. K. and Goldenberg-Hart, D. (2014) "Digital Scholarship Centers: Trends and Good Practices." Coalition for Networked Information. Available online at https://www.cni.org/wp-content/uploads/2014/11/CNI-Digitial-Schol.-Centers-report-2014.web_.pdf.

EGOlink: Supporting Editors of Online Historical Sources through Automatic Link Discovery

Hatem Mousselly Sergieh
mousselly-sergieh@ukp.informatik.tu-darmstadt.de
UKP Lab, Technische Universität Darmstadt, Germany

Michael Piotrowski
Leibniz Institute of European History (IEG), Germany
piotrowski@ieg-mainz.de

Iryna Gurevych
gurevych@ukp.informatik.tu-darmstadt.de
UKP Lab, Technische Universität Darmstadt, Germany

Introduction

EGO (European History Online) is a transcultural history of Europe on the Internet published by the Leibniz Institute of European History. The success and the consequential growth of EGO is, however, a challenge for the editorial office. The interlinking of EGO articles with each other and with external resources is an important aspect of EGO's conceptual design, so each new article has to be integrated into the existing link structure. This means not only that the new article has to be linked to relevant existing articles or online sources, but the copy editors must also check whether links to the new article need to be added to existing articles.

Doing the linking manually is a tedious task. Therefore, we aim in the EGOlink project at developing methods for (semi-)automatically linking EGO articles in order to reduce the manual effort for the editorial office and to improve the navigation for readers.

As a first step, we present a tool for visualizing and analyzing the current link structure in EGO document collection. Besides detecting problems such as under-linked articles, the analysis tool provides information that is needed by a (semi-) automatic linking system. In the next step, the automatic generation of links is addressed; the two main research questions here are: (1) the automatic identification and ranking of potential link anchors in EGO articles, and (2) the discovery of suitable link targets in other EGO articles or external resources.

Background

EGO Document Collection

EGO provides a platform for publishing refereed academic articles about the history of Europe with a special focus on modern Europe from the end of the Middle Ages up to contemporary history. EGO is bilingual, i.e., it contains articles in English and German. Most articles written in one language are translated into the other language and published in both languages. Currently, EGO contains 670 articles: 436 in German and 234 in English.

Linking articles to each other as well as to external resources is a major focus of EGO. Once a new article is submitted the editorial staff put a special effort on linking the articles to already existing EGO articles as well as to other resources or multi-media contents, such as images, interactive maps, video and audio clips.

Entity Linking

As already mentioned, linking articles is at the heart of EGO. So far, this task has been done manually. We aim to optimize this process by developing a (semi-)automatic linking tool based on state-of-the-art approaches while taking the peculiarities of EGO into account.

The problem of (semi-)automatically linking EGO articles to each other or to external resources can be modeled as link discovery, or entity linking (Larson, 2010). In general, entity linking consists of two main steps (Erbs et al, 2011): 1) anchor discovery, which identifies text mentions that can function as link anchors, and 2) target discovery which identifies the best matching references for each anchor from a set of candidates. While anchor discovery can be handled independently of the document collection at hand – e.g. using NER (Named Entity Recognition) (Nadeau et al, 2007) or keyphrase extraction methods (Turney, 2000), discovering and ranking the targets are much affected by the nature of the document collection.

Currently, several state-of-the-art algorithms for entity linking have been made available (Hoffart et al, 2011) (Ceccarelli et al, 2013). However, applying such algorithms directly to perform entity linking based on EGO is challenging. In general, current algorithms rely on extracting statistics about candidate entities, e.g., popularity-based priors; form a huge corpus like Wikipedia. The priors are then used during the process of the target discovery and ranking. However, such an approach is not applicable in our case due to the tiny size of EGO compared to other similar resources like Wikipedia. Another challenge is posed by the German part of EGO. Current approaches focus on English and a special effort should be made to address entity linking for German.

Erbs et al (2011) demonstrated that approaches for target ranking that leverage the linking structure available in the document collection perform best compared to other methods that only leverage the titles of the documents or the associated texts. In light of this result, we propose a tool for analyzing the linking structure in EGO's document collection. The proposed tool helps us to understand the current status of EGO's links and to gain insight regarding the decisions that should be taken to develop the actual entity linking system. Besides identifying links among the articles, the proposed tool also extracts information from an additional resource, namely the literature referenced by the articles. Our hypothesis is that articles sharing the same reference should be linked to each other. As a result, this kind of analysis can help the (semi-) automatic linking system to identify further anchors, thus the interlinking density of the document collection can be increased.

EGO Analyzer

EGO Analyzer is a web-based analysis tool for EGO document collection. The tool can handle the English as well as the German parts of EGO jointly or separately. Since EGO's articles exist in XML format, the tool was built to work on XML data directly and efficiently using the high-performance and scalable XML Database engine BaseX. Besides providing different kinds of metadata about EGO's articles, the tool allows analyzing EGO from different perspectives: the corpus, the article and the reference perspectives (Figure 1).

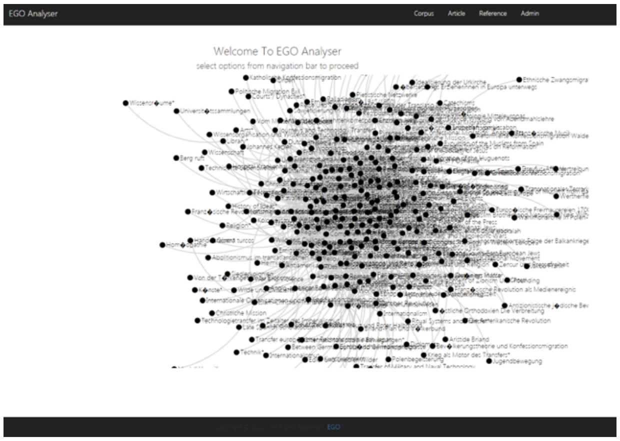
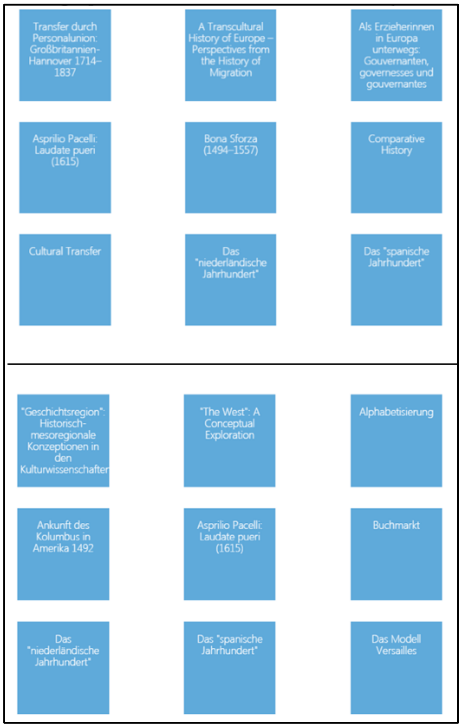


Figure 1. Screenshot: EGO Analyzer

For instance, given the article titled “Kulturtransfer” from the German collection, the tool extracts links to other articles in both directions, i.e., outgoing as well as incoming links (Figure 2). Furthermore, we can view the set of articles that reference the same bibliographical entries as the source article (Table 1).



Transfer durch Personalunion: Großbritannien-Hannover 1714–1837

A Transcultural History of Europe – Perspectives from the History of Migration

Als Erzieherinnen in Europa unterwegs: Gouvernanten, governesses und gouvernantes

Asprilio Pacelli: Laudate pueri (1615)

Bona Sforza (1494–1557)

Comparative History

Cultural Transfer

Das "niederländische Jahrhundert"

Das "spanische Jahrhundert"

"Geschichtsregion": Historisch-mesoregionale Konzeptionen in den Kulturwissenschaften

"The West": A Conceptual Exploration

Alphabetisierung

Ankunft des Kolumbus in Amerika 1492

Asprilio Pacelli: Laudate pueri (1615)

Buchmarkt

Das "niederländische Jahrhundert"

Das "spanische Jahrhundert"

Das Modell Versailles

Figure 2. Source article: “Kulturtransfer”. Top: articles referred to by the source article (outgoing links). Bottom: articles referring to the source article (incoming links)

Article	Common Bibliography
Cultural Transfer	5
Eine transkulturelle Geschichte Europas – migrationsgeschichtliche Perspektiven	2
Europäisierungen	2

Table 1. Articles sharing bibliographical entries with the source article “Kulturtransfer”

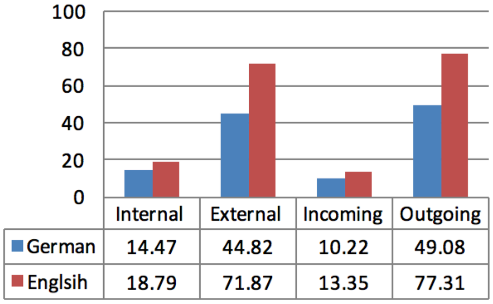
Results of the Analysis

Using EGO Analyzer we collected statistics about the linking structure in EGO. We divided the links into two categories: internal and external. Internal links point from one EGO article to another, while external links refer to external targets. For each article, we also distinguish between two types of links: incoming and outgoing links. Figure 3 shows the average number of internal/external as well as incoming/outgoing links per article.

The results show that dominance of the external links (3 times as much as internal links). This can be explained due to the relatively small size of the document collection. Accordingly, it is hard for the editors to find corresponding internal links for entities in a newly added article. Moreover, this also explains the higher number of outgoing links per article. Indeed, most of these links go to external resources which are not necessarily linked back to EGO articles. This observation demonstrates the urgent need for an automatic linking system that enables the editors to increase the interconnectivity of EGO.

Furthermore, the linking tool can also run as a background process that updates the linking structure of EGO every time a new article is added.

Regarding the shared bibliography, we identified 13,309 unique references, 737 of them are referenced by at least two articles. We will use this kind of information as auxiliary input for the linking algorithm to identify additional linking candidates.



	Internal	External	Incoming	Outgoing
🟦 German	14.47	44.82	10.22	49.08
🟥 Englsih	18.79	71.87	13.35	77.31

Figure 3: Average number of links per article and category

Conclusion

In this abstract, we presented work in progress for assisting editors of EGO to (semi-)automatically link the articles. We focused on the first step towards such a system, namely, analyzing the link structure in the target document collection. We presented a web-based tool that is able to perform such analysis. Currently, we are working on the actual linking system as well as a user-friendly interface for the editorial office. By the time of the conference, we will also provide a demonstration of the actual linking tool.

Bibliography

Ceccarelli, D., et al. (2013). Dexter: an open source framework for entity linking. Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval. 2013, pp. 17--20.

Erbs, N., Zesch, T., and Gurevych, I. (2011). Link discovery: A comprehensive analysis. Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. 2011, pp. 83--86

Hoffart, J, et al. (2011). Robust disambiguation of named entities in text. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011, pp. 782--792.

Larson, R. R. (2010). Information retrieval: Searching in the 21st century; human information retrieval. s.l. : Wiley Online Library.

Nadeau, D., and Sekine, S. (2007). A survey of named entity recognition and classification. 2007, Vol. 30, 1, pp. 3--26.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. 2000, Vol. 2, pp. 303--336.

Seven Years of Research on Digital 3D Reconstruction in Humanities – Results, Implications and Perspectives

Sander Münster
sander.muenster@tu-dresden.de
Technische Universität Dresden, Germany

Abstract

Since an academic discourse on using digital 3D reconstruction tools for humanities research is still highly application-oriented, an overarching objective of our research work is to draw a currently missing “big picture” on digital 3D reconstruction as a research tool in digital humanities by combining theoretically and practically grounded research parts as well as including multi-disciplinary perspectives. This article provides an overview of research work carried out by our department in 12 projects over the past seven years, namely on (a) scenarios and practices for the employment of digital 3D reconstruction and visualization methods and approaches for scientific research in visual humanities, (b) requirements and recommendations for digital tools, and (c) approaches to teach digital 3D reconstruction methods at university. The article is intended to contribute to a comprehensive discourse on a unique epistemology of digital 3D reconstruction in humanities and to deliver related practically tested implications for software de-sign, teaching and organizational settings.

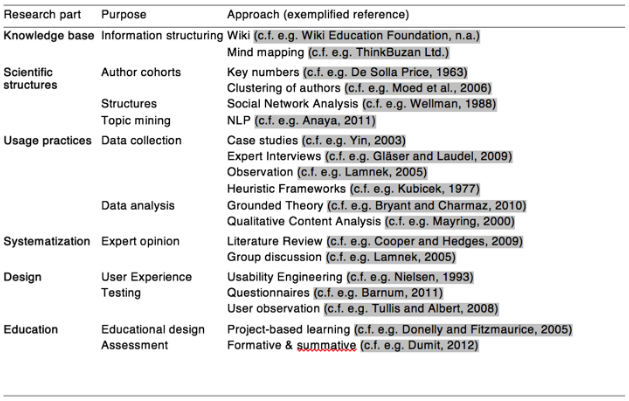
Introduction

For more than 30 years, digital 3D modelling and in particular reconstruction methods have been widely used to support research and education in the digital humanities, especially but not exclusively on historical architecture. While technological backgrounds, project opportunities, and methodological considerations for the ap-plication of digital 3D reconstruction tech-niques are widely discussed in literature (e.g. Arnold and Geser, 2008, European Commission, 2011, Frischer, 2008, Bentkowska-Kafel et al., 2012, Bentkowska-Kafel, 2013, Kohle, 2013), our interest is to investigate digital 3D reconstruction as a scholarly area and to derive implications for further organizational and methodical development. Against this background, this research work is dedicated to support the dissemination process by investigating the following research questions:

- What scenarios and practices exist for the employment of digital methods and approaches for scientific research within the field of digital 3D reconstruction, especially in art and architectural history?
- What are requirements and recommendations for the design of tools and media related to digital 3D re-construction?
- How can the use of digital methods and, in particular, digital 3D reconstruction techniques in visual human-ities be learned and taught?

Research

Against the background of these research interests, our research activities include to investigate (1) a scholarly community, (2) usage practices occurring with-in single projects and to gain implications for an appropriate organizational development. Moreover, practical application and a development of implications for (4) the conception and creation of a user-centered design of software tools and environments for digital humanities scholars and an (5) implementation in student education are in focus. Research methods and approaches used within the described research activity mainly derive from social and information sciences as well as from science and technology studies (cf. Hackett et al., 2008, Boczkowski and Lievrouw, 2008, pp. 955) and comprise e.g. bibliometric analysis (cf. Vinkler, 1996), quantitative and qualitative empirical analysis as well as structuring approaches like mind mapping (Table 1). Relevant research has been active since 2010 in 12 projects (ongoing till approx. 2020) on the local, national and EU levels, with our department’s participation.



Research part	Purpose	Approach (exemplified reference)
Knowledge base	Information structuring	Wiki (c.f. e.g. Wiki Education Foundation, n.a.) Mind mapping (c.f. e.g. ThinkBuzan Ltd.)
Scientific structures	Author cohorts	Key numbers (c.f. e.g. De Solla Price, 1963) Clustering of authors (c.f. e.g. Moed et al., 2006)
	Structures	Social Network Analysis (c.f. e.g. Wellman, 1988)
	Topic mining	NLP (c.f. e.g. Anaya, 2011)
Usage practices	Data collection	Case studies (c.f. e.g. Yin, 2003) Expert Interviews (c.f. e.g. Gläser and Laudel, 2009) Observation (c.f. e.g. Lamnek, 2005) Heuristic Frameworks (c.f. e.g. Kubicek, 1977)
	Data analysis	Grounded Theory (c.f. e.g. Bryant and Charmaz, 2010) Qualitative Content Analysis (c.f. e.g. Mayring, 2000)
Systematization	Expert opinion	Literature Review (c.f. e.g. Cooper and Hedges, 2009) Group discussion (c.f. e.g. Lamnek, 2005)
Design	User Experience Testing	Usability Engineering (c.f. e.g. Nielsen, 1993) Questionnaires (c.f. e.g. Barnum, 2011) User observation (c.f. e.g. Tullis and Albert, 2008)
Education	Educational design	Project-based learning (c.f. e.g. Donelly and Fitzmaurice, 2005)
	Assessment	Formative & summative (c.f. e.g. Dumit, 2012)

Table 1. Brief overview of methodical approaches

Results

Considering a scholarly community on digital 3D reconstruction and modeling, discourses on major conferences during the last 25 years were mainly led by institutions from European Mediterranean countries, covering primarily technological topics. In particular, statues and buildings in Mediterranean countries dating from all periods Anno Domini deliver rich content for such reconstruction (Fig. 1). Institutions with high numbers of publications, connections to co-authors or the extraordinary importance to link groups of researchers to each other can be identified on a structural level (cf. Fig. 2). Due to the high complexity and team based workflows, aspects and usage practices for communication, cooperation, and quality management are of high relevance within 3D reconstruction projects. Especially if people with different disciplinary backgrounds are involved, visual media are intensively used to foster communication and quality negotiations (cf. Fig. 3), for example by comparing source images and renderings of the created virtual reconstruction. Furthermore, several projects successfully adopted highly standardized conventions from architectural drawings for interdisciplinary exchange (Münster, 2013). To support an organizational development, we ran five workshops to identify ongoing research topics and challenges, involving around 100 researchers in total. Current challenges named (cf. Fig. 4) aim at a research and development of sustainable and practicable approaches to access wider scientific communities and audiences and include aspects such as widely interoperable documentation and classification strategies and schemes, an overarching cataloguing of projects, and the creation of objects as well as strategies and technologies for an exchange between different technological domains and approaches of usage. Regarding design implications for digital environments, we investigated, for example, that relatively little visual information is needed to allow observers to distinguish buildings from each other or to identify a single building and gain information about its spatial relation and shape (Münster et al., 2017b). Moreover, we adopted and evaluated team project-based learning approaches to support student education in digital 3D reconstruction (cf. Fig. 6). As observed in two courses so far, a development of procedures and strategies for cooperation within student project teams for creating virtual representations evolves slowly and is mostly caused by emerging problems and urgent demands. Related competencies are based highly on implicit knowledge and experience. As a consequence, a teaching of implications and best practices prior to commencing a project is less effective than coaching during the project work.

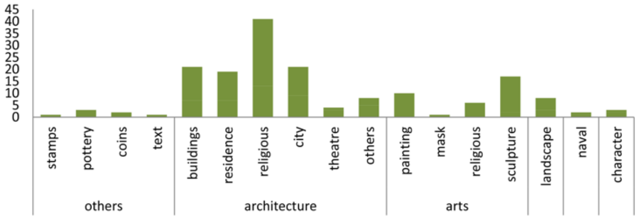


Figure 1. Types of 3D reconstructed and modelled artifacts (Sample: 478 publications on digital 3D modeling in humanities)

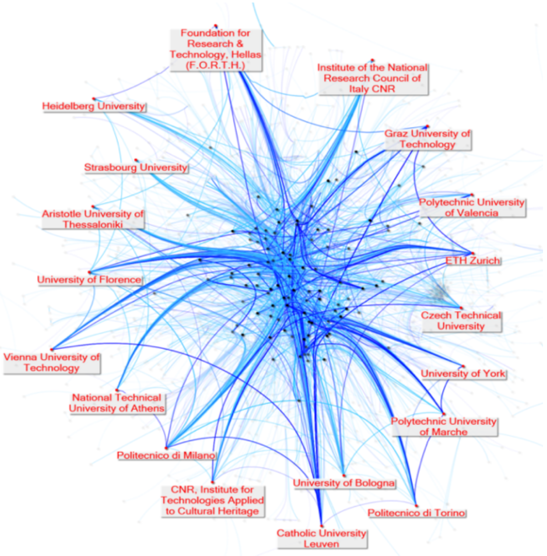

Foundation for Research & Technology, Hellas (F.O.R.T.H.)
Institute of the National Research Council of Italy CNR
Heidelberg University
Graz University of Technology
Strasbourg University
Polytechnic University of Valencia
Aristotle University of Thessaloniki
ETH Zurich
University of Florence
Czech Technical University
Vienna University of Technology
University of York
National Technical University of Athens
Polytechnic University of Marche
Politecnico di Milano
CNR, Institute for Technologies Applied to Cultural Heritage
University of Bologna
Politecnico di Torino
Catholic University Leuven

Figure 2. Accumulated Author-Co-Author Relations by institutions (TOP 10 in Degree, Count, Betweenness Centrality named, Sample: 3917 publications on digital 3D modeling in humanities)

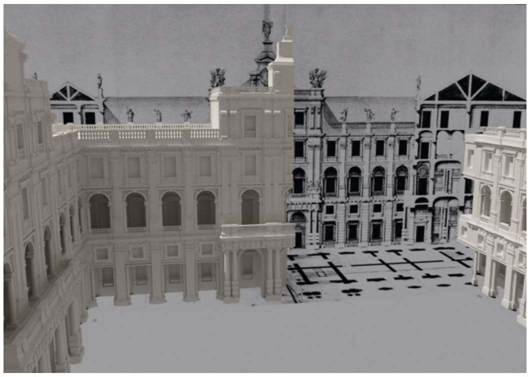
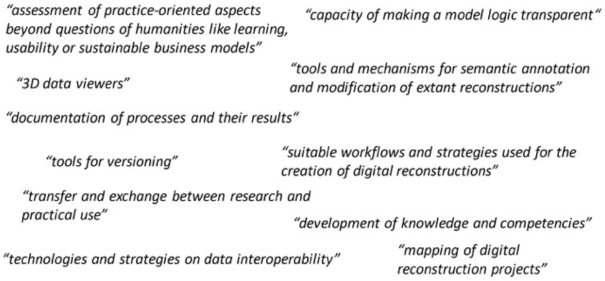


Figure 3. Exemplified modeling as well as quality control techniques. Originally published in (Münster et al., 2017a).



"assessment of practice-oriented aspects beyond questions of humanities like learning, usability or sustainable business models"

"capacity of making a model logic transparent"

"3D data viewers"

"tools and mechanisms for semantic annotation and modification of extant reconstructions"

"documentation of processes and their results"

"tools for versioning"

"suitable workflows and strategies used for the creation of digital reconstructions"

"transfer and exchange between research and practical use"

"development of knowledge and competencies"

"technologies and strategies on data interoperability"

"mapping of digital reconstruction projects"

Figure 4. (Selected) future challenges named for digital 3D reconstruction in academic contexts (outcome from a workshop series involving around 60 international researchers).

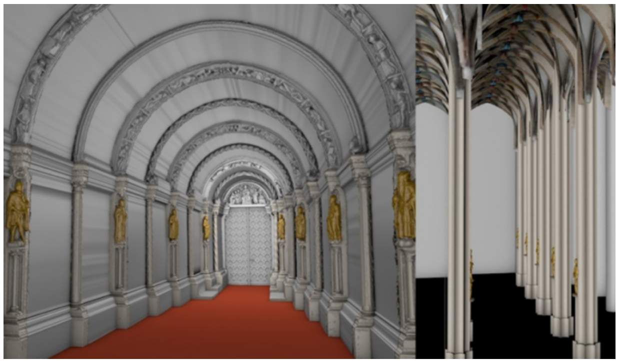


Figure 5. Results from Freiberg Cathedral App educational project. Models: Wachsmuth et al., 2014

Summary

Since it is our vision to establish digital 3D reconstruction as a scholarly accepted and widely used research method in humanities, it seems to be crucial to add a critically reflected methodological basis and anchor it in academic culture. To draw this currently missing "big picture", our department performed research on digital 3D reconstruction within various projects by combining both theoretically and practically grounded research parts. This comprises research on scholarly communities, scholarly practices and methodological recommendations as well as implications for interaction design, teaching and dissemination.

Bibliography

Anaya, L. H. (2011). *Comparing Latent Dirichlet Allocation And Latent Semantic Analysis As Classifiers,* Denton, University Of North Texas.

Arnold, D. & Geser, G. (2008). *Epoch Research Agenda – Final Report*, Brighton.

Barnum, C. M. (2011). *Usability Testing Essentials : Ready, Set...Test!,* Burlington, Morgan Kaufmann.

Bentkowska-Kafel, A. (2013). *Mapping Digital Art History,* Los Angeles, Getty Conservation.

Bentkowska-Kafel, A., Denard, H. & Baker, D. (2012). *Paradata And Transparency In Virtual Heritage,* Burlington, Ashgate.

Boczkowski, P. & Lievrouw, L. A. (2008). Bridging Sts And Communication Studies. In: Hackett, E. J., Amsterdamska, O., Lynch, M. & Wajcman, J. (Eds.) *The Handbook Of Science And Technology Studies.* 3rd Ed. Ed. Cambridge: Mit Press.

Bryant, A. & Charmaz, K. (2010). *The Sage Handbook Of Grounded Theory,* Thousand Oaks, Sage.

Cooper, H. & Hedges, L. V. (2009). *The Handbook Of Research Synthesis And Meta-Analysis*, New York, Russell Sage Foundation.

De Solla Price, D. (1963). *Little Science - Big Science,* New York, Columbia Univ. Press.

Donelly, R. & Fitzmaurice, M. (2005). Collaborative Project-Based Learning And Problem-Based Learning In Higher Education: A Consideration Of Tutor And Student Role In Learner-Focused Strategies. In: O'neill, G., Moore, S. & Mcmullin, B. (Eds.) *Emerging Issues In The Practice Of University Learning And Teaching. Dublin: All Ireland Society For Higher Education* (Aishe)

Dumit, N. Y. (2012). Diagnostic/Formative/Summative Assessment, N.N.

European Commission (2011). Survey And Outcomes Of Cultural Heritage Research Projects Supported In *The Context Of Eu Environmental Research Programmes. From 5th To 7th Framework Programme,* Brussels, European Commision.

Frischer, B. (2008). Beyond Illustration : 2d And 3d Digital Technologies As Tools For Discovery In Archaeology, Oxford, Tempus Reparatum.

Gläser, J. & Laudel, G. (2009). *Experteninterviews Und Qualitative Inhaltsanalyse Als Instrumente Rekonstruierender Untersuchungen*, Wiesbaden, Vs Verlag Für Sozialwissenschaften.

Hackett, E. J., Amsterdamska, O., Lynch, M. & Wajcman, J. (2008). The Handbook Of Science And Technology Studies, Cambridge, Mit Press.

Kohle, H. (2013). *Digitale Bildwissenschaft,* Glückstadt.

Kubicek, H. (1977). Heuristische Bezugsrahmen Und Heuristisch Angelegte Forschungsdesigns Als Element Einer Konstruktionsstrategie Empirischer Forschung. In: Köhler, R. (Ed.) *Empirische Und Handlungstheoretische Forschungskonzeptionen In Der Betriebswirtschaftslehre*. Stuttgart.

Lamnek, S. (2005). Qualitative Sozialforschung. Lehrbuch, Weinheim.

Mayring, P. (2000). Qualitative Content Analysis. Forum Qualitative Sozialforschung, 1, Art. 20.

Moed, H. F., Glänzel, W. & Schmoch, U. (2006). *Handbook Of Quantitative Science And Technology Research: The Use Of Publication And Patent Statistics In Studies Of S&T Systems,* Springer Science & Business Media.

Münster, S. (2013). Workflows And The Role Of Images For A Virtual 3d Reconstruction Of No Longer Extant Historic Objects. Ann. Photogramm. *Remote Sens. Spatial Inf. Sci.*, Ii-5/W1 (Xxiv International Cipa Symposium), 197–202.

Münster, S., Jahn, P.-H. & Wacker, M. (2017a.) Von Plan- Und Bildquellen Zum Virtuellen Gebäudemodell. Zur Bedeutung Der Bildlichkeit Für Die Digitale 3d-Rekonstruktion Historischer Architektur. In: Ammon, S. & Hinterwaldner, I. (Eds.) *Bildlichkeit Im Zeitalter Der Modellierung. Operative Artefakte In Entwurfsprozessen Der Architektur Und Des Ingenieurwesens*. München: Wilhelm Fink Verlag.

Münster, S., Kröber, C., Weller, H. & Prechtel, N. (2017b). Virtual Reconstructions Of Historical Architecture As Media For Visual Knowledge Representation. In: Ioannides, M., Magnenat-Thalmann, N. & Papagiannakis, G. (Eds.) *Mixed Reality And Gamification For Cultural Heritage.* Cham: Springer Lncs.

Nielsen, J. (1993). *Usability Engineering,* Salt Lake City, Academic Press.

Thinkbuzan Ltd. (n.d.) Mind Mapping: Scientific Research And Studies.

Tullis, T. & Albert, B. (2008). *Measuring The User Experience,* Amsterdam, Mrgan Kaufman Publishers.

Vinkler, P. (1996). Some Practical Aspects Of The Standardization Of Scientometric Indicators. *Scientometrics*, 35, 237-245.

Wellman, B. (1988). Structural Analysis. From Method And Metaphor To Theory And Substance. In: Wellman, B. & Berkowitz, S. D. (Eds.) *Social Structures: A Network Approach.* Princeton: Princeton University Press.

Wiki Education Foundation (n.d.) *N.A. Theories: Wikipedia And The Production Of Knowledge.*
Yin, R. K. (2003). *Applications Of Case Study Research,* Thousand Oaks, Sage.

A Collaborative Approach between Art History and Literature via IIIF

Kiyonori Nagasaki
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

Tetsuei Tsuda
tsuda@tobunken.go.jp
Tokyo Research Institute for Cultural Properties, Japan

X. Jie Yang
xyang@ucalgary.ca
University of Calgary, Canada

Yuho Kitazaki
yuho.kitazaki@gmail.com
University of Tokyo, Japan

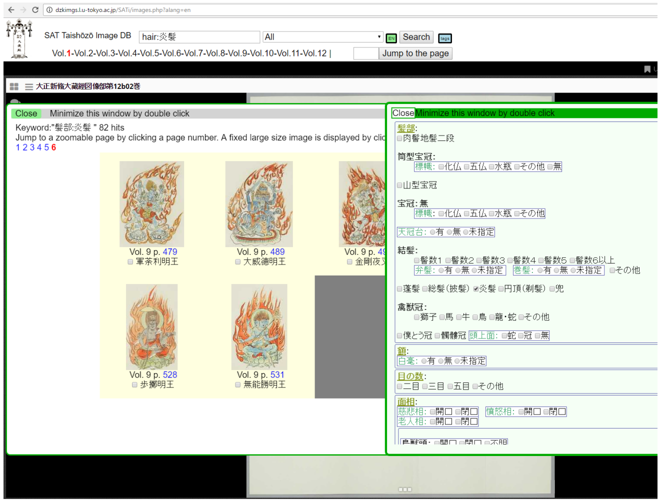
A. Charles Muller
acmuller@l.u-tokyo.ac.jp
Univerisity of Tokyo, Japan

Masahiro Shimoda
shimoda@l.u-tokyo.ac.jp
University of Tokyo, Japan

Introduction

This paper describes a collaborative approach of information exchange between art history and literature via IIIF as conducted by two projects: the SAT Daizōkyō Text Database Committee (SAT) and a project to leverage an open dataset of the National Institute of Japanese Literature. This approach is technically similar to that used by previous projects such as TILE and TEI. However, as it has not been easy for them to fully treat the binding of images of a book between Web services, this approach adopts the IIIF (International Image Interoperability Framework) so that both can connect easily and efficiently. After explaining the distinctive aspects of both projects, we will introduce a collaborative solution.

In May 2016, SAT released the SAT Taishōzō Image DB (SATiDB), which includes digital facsimiles of a series of Buddhist images and their interpretations in the Taishō Tripitaka consisting of 12 volumes originally published in 1933. SATiDB provides annotations for about 5,200 Buddhist icons (*busson*) and symbols (*sanmayagyō* and mandala) in the books and several search functions of the annotations with a simple translator from English to technical terms in CJK characters via the Digital Dictionary of Buddhism. As the annotations use a vocabulary of attributes of Buddhist icons, such as hair style, sitting style, type of chair, possessions, etc., which was defined by this project due to the absence of such a vocabulary in the source data, users can also search the objects by clicking a checkbox of one or more term in a list form of the vocabulary (Figure 1).



dzkimgs.l.u-tokyo.ac.jp/SATi/images.php?alang=en

SAT Taishōzō Image DB hair:炎髪 All EN Search tags

Vol.1-Vol.2-Vol.3-Vol.4-Vol.5-Vol.6-Vol.7-Vol.8-Vol.9-Vol.10-Vol.11-Vol.12 | Jump to the page

大正新脩大藏經図像部第12b02巻

Close Minimize this window by double click

Keyword:"髪部:炎髪 " 82 hits

Jump to a zoomable page by clicking a page number. A fixed large size image is displayed by clic

1 2 3 4 5 6

Vol. 9 p. 479 □ 軍荼利明王
Vol. 9 p. 489 □ 大威德明王
Vol. 9 p. 49 □ 金剛夜叉
Vol. 9 p. 528 □ 歩擲明王
Vol. 9 p. 531 □ 無能勝明王

Close Minimize this window by double click

髪部:
□肉髻地髪二段

筒型宝冠:
標幟: □化仏 □五仏 □水瓶 □その他 □無

□山型宝冠

宝冠: 無
標幟: □化仏 □五仏 □水瓶 □その他

天冠台: ○有 ○無 ○未指定

結髪:
□髻数1 □髻数2 □髻数3 □髻数4 □髻数5 □髻数6以上
弁髪: ○有 ○無 ○未指定 巻髪: ○有 ○無 ○未指定 □その他

□蓬髪 □総髪(披髪) ☑炎髪 □円頂(剃髪) □兜

禽獣冠:
□獅子 □馬 □牛 □鳥 □龍・蛇 □その他

□僕とう冠 □髑髏冠 頭上面: □蛇 □冠 □無

額:
白毫: ○有 ○無 ○未指定

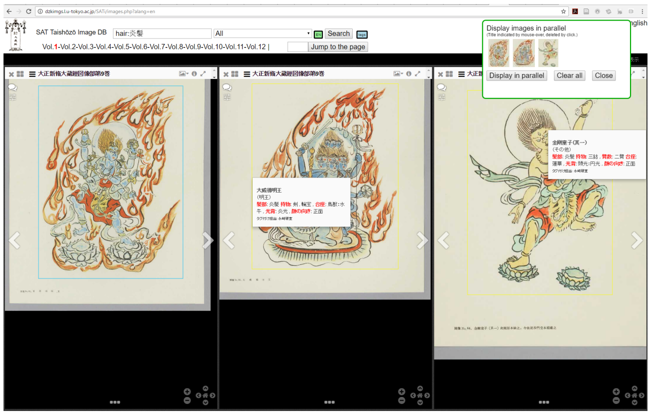
目の数:
□二目 □三目 □五目 □その他

面相:
慈悲相: □開口 □閉口 憤怒相: □開口 □閉口
老人相: □開口 □閉口

鳥獣頭: □開口 □閉口 □不明

Figure 1. A search result of “burning hair”

As the system is compliant with IIIF, the images and annotations can be leveraged in various ways, even from other web sites. SATiDB has a function to expose several objects in parallel by clicking checkboxes of cropped images by coordination of each object in the search results on the IIIF viewer, Mirador (Figure 2).



dzkimgs.l.u-tokyo.ac.jp/SATi/images.php?alang=en

SAT Taishōzō Image DB hair:炎髪 All EN Search tags

Vol.1-Vol.2-Vol.3-Vol.4-Vol.5-Vol.6-Vol.7-Vol.8-Vol.9-Vol.10-Vol.11-Vol.12 | Jump to the page

Display images in parallel
(Title indicated by mouse-over, deleted by click.)
Display in parallel Clear all Close

大正新脩大藏經図像部第9巻

大威德明王
(明王)
髪部: 炎髪 持物: 剣，輪宝，台座: 鳥獣:水牛，光背: 炎光，顔の向き: 正面
タグ付け担当: 永崎研宣

金剛童子(其一)
(その他)
髪部: 炎髪 持物: 三鈷，臂数: 二臂 台座: 蓮華，光背: 頭光:円光，顔の向き: 正面
タグ付け担当: 永崎研宣

Figure 2. Checked images are displayed in parallel.

The annotations were embedded by forty-three researchers in the field of Japanese art history on a web collaboration system in 2015. We developed the system utilizing RedHat Linux, Apache, PostgreSQL, PHP, jQuery, and Annotorious, which enabled the easy annotation of images. The annotations were stored in PostgreSQL including attributes such as date and responsibility. After input, the data were converted into IIIF Presentation API and distributed with hi-resolution images converted from 60M-pixel images delivered with IIIF Image API. This system provides researchers of Japanese art history with a brand new function to see and compare Buddhist icons and symbols. Many positive comments have been received from the researchers of Japanese art history and the number of visitors of the site was over 12,000 in the first month, but no papers have yet been produced explicitly using this system.

The other project also developed a Web collaboration system to embed transcription of Japanese texts (the issues of such transcription have been described by Nagasaki et al, 2016) line-by-line in the style of IIIF annotation which enables to search images as-they-are via Smart-GS. (Hashimoto et al, 2014) It adopts OpenSeaDragon and its plugins to annotate images with zooming and has a function to convert them into the format of IIIF Presentation API. So far, two pre-modern woodcut printing books written in cursive Japanese script were completely done by two researchers and available via customized Mirador for right-to-left viewing-direction and vertical texts (Figure 3).

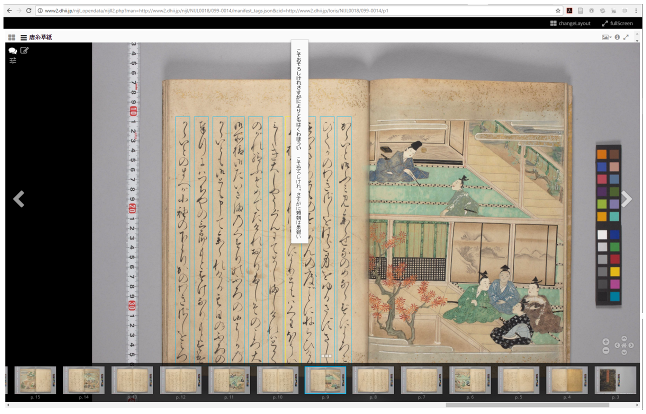
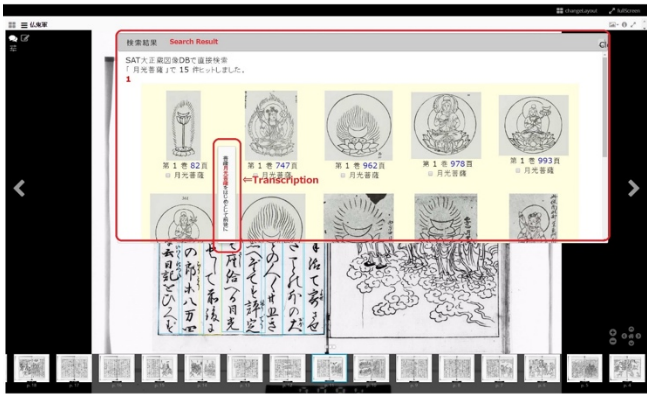


Figure 3. Customized Mirador for right-to-left viewing-direction and vertical texts

Finally, we explain the approach of linking both projects. As one of the two transcribed woodcut printing books includes names of Buddhist saints, we added tags on the names to trigger an event to search the name and prepare a function to request queries to the SATiDB. On the other hand, in the SATiDB, a function to distribute only a list of search results including images cropped by IIIF Image API was implemented to pull search results from other Web sites by use of a form of URL such as:
http://dzkimgs.l.u-tokyo.ac.jp/SATi/key:_*keyword*_



SAT大正蔵図像DBで直接検索
「 月光菩薩 」で 15 件ヒットしました。
1

Figure 4. Search result of SATiDB by clicking a red-colored part of transcribed text

As a result, readers– primarily researchers, but laypersons as well– can see images of related Buddhist icons on SATiDB while reading the book. See Figure 4 to understand background of it. This is a typical solution of IIIF and easily applicable for any environment in the digital humanities.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP15H05725, JP16H03422.

Bibliography

Hashimoto, Y., Aihara, K., Hayashi, S., Kukita, M., Ohura, M., (2014)The SMART-GS Project: An Approach to Image-based Digital Humanities, Digital Humanities Conference 2014, http://dharchive.org/paper/DH2014/Poster-48.xml.

Nagasaki, K., Tomabechi, T., Muller, A. C., Shimoda, M.(2016) “Digital Humanities in Cultural Areas Using Texts That Lack Word Spacing”, Digital Humanities 2016, http://dh2016.adho.org/abstracts/416

Building Entity-Centric Event Collections For Supporting Research in Political and Social History

Federico Nanni
federico@informatik.uni-mannheim.de
University of Mannheim, Germany

Nikolay Marinov
marinov@sowi.uni-mannheim.de
University of Mannheim, Germany

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de
University of Mannheim, Germany

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire, United States of America

Introduction

The World Wide Web provides the research community with an unprecedented abundance of primary sources for diachronically tracing, examining and understanding major events and transformations in our society (such as the rise of euroscepticism or the impact of the recent economic crisis). For two decades, public and private institutions have preserved these born-digital materials for future analysis (Gomes and Costa, 2011). However, these collections are now so large that it is infeasible for researchers to study political and social phenomena by examining them in their entirety.

Creating event collections. A common solution that web archives are currently adopting for sustaining the use of the collected sources in humanities research is to offer topic-specific collections. For example, on Archive-it, the Internet Archive presents a few collections on large-scale events such as the Boston Marathon Bombing, Black Lives Matter and the Charlie Hebdo terrorist attack. The collections are curated "by the Archive-It team in conjunction with curators and subject matter experts from institutions around the world" .

Another solution for creating event collections from large datasets is a filtering approach that collects only documents that mention the name of the event; this method has been employed for example in temporal summarization tasks (see Aslam et al., 2013).

Current limitations. The collections created following one of these two approaches share crucial limitations: a) they are small in number; b) the selection process is not always transparent; c) they generally offer only documents that are closely related to an event but lack information on background stories as well as contextual clues. Especially the latter is a crucial issue for historical analyses.

Our vision. We are currently developing a solution for creating event collections that identifies not only the core documents related to the event itself, but most importantly sub-groups of documents which describe related aspects. We do so through an expansion process that is informed by latently relevant concepts and entities from a knowledge base, whose presence in documents is interpreted as one of many indicators of relevance.

Specific contribution. At the DH conference we intend to present the final results of our study, together with its application for supporting research in political and social history.

Method

Let us consider an event, for example the Syrian Civil War, as a node in a knowledge graph (e.g. DBpedia).

As a first step, a domain expert will identify a series of other entities in the knowledge graph that are highly related to the event (in Nanni et al., 2016, we show that this step could be automatized adopting a simple relatedness measure). These could be people, such as Bashar Al-Assad as well as countries (e.g. Turkey, Russia, United States), concepts (e.g. Protests) and other specific events (e.g. The Refugee Crisis). These initial seeds will support us in retrieving other related entities and concepts from the knowledge graph in an automated fashion (we described our solution in Nanni et al. 2016).

While retrieving related entities is important, these are meaningless without human-readable descriptions of the entity's relation to the event. As a matter of fact, the entity United States has many different aspects, and only few of them are related to the event Syrian Civil War.

In order to retrieve entities in context, we use Wikipedia as an initial corpus. Next, relevant passages from the documents are identified in the collection by information retrieval. Having the entity in context will tell us with which words, concepts and other entities it frequently appears together (a complete overview of the method is presented in Nanni et al., 2017). For example, if a document mentions the United States together with James Foley and ISIS, it is likely to be related to the Syrian Civil War, even without mentioning these words explicitly.

Case studies

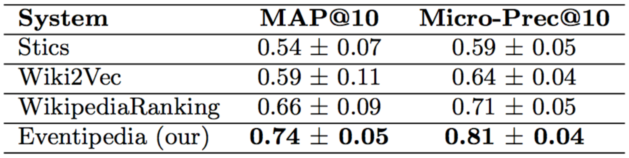
We are currently working on two different research tasks:

1. The first study is focused on identifying political speeches on foreign events (such as elections in other countries) in the US Congressional Records (1989-2016), which are available on Congress.gov and through the Internet Archive (Congress.gov provides full-text access to daily congressional record issues dating from 1995, beginning with the 104th Congress. Proceedings for previous years are available on THOMAS). The goal is to measure the amount of attention that US politicians give to international events in correlation with other internal affairs.

2. In the second study, we intend to detect similar patterns during the early rise of anti-establishment protests. Our aim is to uncover small events, which did not turn into large-scale insurrections and therefore are not studied sufficiently. The work is conducted on a large (16 terabyte) web archive of news, blogs, forums and social media, namely the TREC Streaming Corpus (This corpus is a huge web archive collection collected between 2012 and 2014). Finally, the goal of the project is to obtain a better understanding on how and why specific protests succeed while others do not (also in correlation with analyses from the previous study).

Experiments

Identifying related entities. In a previous work (Nanni et al., 2016), we have first established the quality of our entity-relatedness solution (Eventipedia), by comparing it with a series of other baselines commonly used in the field. The results are reported in Figure 1.

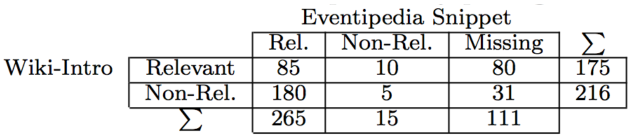


System	MAP@10	Micro-Prec@10
Stics	0.54 ± 0.07	0.59 ± 0.05
Wiki2Vec	0.59 ± 0.11	0.64 ± 0.04
WikipediaRanking	0.66 ± 0.09	0.71 ± 0.05
Eventipedia (our)	**0.74** ± **0.05**	**0.81** ± **0.04**

Figure 1. Evaluation on entity-event relatedness (from Nanni et al., 2016).

While our entity-relatedness solution outperformed the other tested methods, it also showed a few limitations. In fact, this approach (in its fully automated fashion) tends to privilege specific entities over the most commonly mentioned entities. We address this in developing techniques that are supervised by domain experts; this ensures us to always consider the most relevant related entities.

Collect entities in context. Additionally we studied the identification of human-readable descriptions of the entity's connection to the event. We compared our entity link-based approach with a common information retrieval heuristic which considers the first sentences of the entity's Wikipedia article, as a relevant passage (Wiki-Intro).



		Eventipedia Snippet			
		Rel.	Non-Rel.	Missing	∑
Wiki-Intro	Relevant	85	10	80	175
	Non-Rel.	180	5	31	216
	∑	265	15	111	

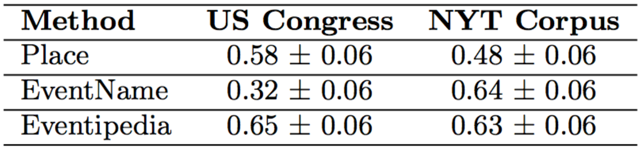
Figure 2. Evaluation on retrieving entities in context (from Nanni et al., 2016).

The results are presented in Figure 2. In 45% of the cases, the Wiki-Intro was a sufficient explanation. However, our Eventipedia approach provides sufficient explanations in 68% of the cases. We remark that for nearly all cases, where Eventipedia does provide a snippet, this is also relevant. In contrast, the Wiki-Intro only provides a good explanation in 42% of the cases. This is because many event-relevant entities (e.g. the United States) are often more popularly known for other accomplishments and therefore the first paragraph is not a good description of entity involvement in the event.

Retrieve relevant documents. We are currently assessing the quality of our information retrieval solution, which uses entities and contextual passages to retrieve documents about specific events with an approach based on Dalton et al. (2014). We report here the very first results of our study on retrieving speeches about foreign elections. This work has been conducted both on the US Congressional Records on the New York Times Corpus.

We compared it with two baselines, a) retrieving all documents that mention the name of the country (e.g. "Syria") and b) retrieving all documents that precisely mention the name of the event (e.g. "election in Syria"). It is evident that the first solution is recall-oriented, while the second, already adopted by the TREC temporal summarization task, favors precision.

Given an event, such as the Syrian presidential election, 2007, all three methods produce a ranking of documents. We examine the quality of the ranking considering 15 different elections. For each election, we created a gold standard of 45 documents (relevant and non relevant). Table 1 presents the results in term of mean-average precision (MAP) on the two datasets.



Method	US Congress	NYT Corpus
Place	0.58 ± 0.06	0.48 ± 0.06
EventName	0.32 ± 0.06	0.64 ± 0.06
Eventipedia	0.65 ± 0.06	0.63 ± 0.06

Table 1. First results (MAP) on the collection-building task.

The results of this initial study lead to a few findings. First we see that, especially on the US Congressional Records, using the name of the event permits to collect a fraction of relevant documents, but not all of them. Second, our solution, which uses related entities in context, provides good performance quality on both datasets. Third, our solution is able to identify materials that do not explicitly mention the name of the event.

"Non-relevant" documents. We next analyze what kind of "non-relevant" documents the different systems retrieved among the top elements of the ranking. Non-relevant documents retrieved by the "Event name" solution often discuss different topics and simply mention the event out of context; these documents could be for example general summaries of the previous week.

Instead, non-relevant documents retrieved by Eventipedia are often related to the political activity of a foreign country, but not specifically about the election. For example, they could mention the visit of a candidate to Washington, a few months before the vote. It is evident that choosing one method over the other will shape the event-collection in a different way. Ultimately, it is up to the humanities researcher to decide which documents are most important for the analysis.

Bibliography

Aslam, J. A., et al. (2013) "TREC 2013 Temporal Summarization." *TREC.*

Dalton, J., Dietz, L., and Allan, J. (2014). "Entity query feature expansion using knowledge base links." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* ACM.

Gomes, D., Miranda, J., and Costa, M. (2011) "A survey on web archiving initiatives." *International Conference on Theory and Practice of Digital Libraries.* Springer Berlin Heidelberg.

Nanni, F., Ponzetto, S. P., and Dietz, L. (2016) “Entity Relatedness for Retrospective Analyses of Global Events”, *Proceedings of NLP+CSS: Workshops on Natural Language Processing and Computational Social Science,* at WebSci.

Nanni, F., Ponzetto, S. P., and Dietz, L. (2017, forthcoming) “Building Entity-Centric Event Collections”, *Proceedings of the Joint Conference on Digital Libraries,* (forthcoming).

Enhancing Domain-Specific Entity Linking in DH

Federico Nanni
federico@informatik.uni-mannheim.de
University of Mannheim, Germany

Yang Zhao
yzhao@mail.uni-mannheim.de
University of Mannheim, Germany

Simone Paolo Ponzetto
simone@informatik.uni-mannheim.de
University of Mannheim, Germany

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire, United States of America

For the purpose of information retrieval and text exploration, digital humanities (DH) scholars have examined the potential of methods such as keyphrase extraction (Hasan and Ng, 2014) and named entity recognition (Nadeau and Sekine, 2007). However, these solutions still face challenges in the presence of polysemy and synonymy (e.g. distinguish between “Paris” the capital of France or the city in Ontario or recognize that “POTUS” and “Barack Obama” might refer to the same person).

Entity Linking

In the last decade, advances in natural language processing (NLP) gave rise to word-sense disambiguation and entity linking techniques (Cornolti et al., 2013), which automatically disambiguate entities and concepts in context and link them to knowledge bases such as Wikipedia, DBpedia (Auer et al., 2007) or Babelnet (Navigli and Ponzetto, 2012). Among them, TagMe (Ferragina and Scaiella, 2010) has been often adopted in NLP, thanks to its decent performance on different datasets and to its easy-to-use API.

Current Limitations for DH research

TagMe also highlights a few common limitations of current standard entity linking systems that reduce their applicability within most scenarios found in the heterogeneous spectrum of Digital Humanities research.

- **Black Box and reproducibility.** As Hasibi et al. (2016) recently remarked, the TagMe RESTful API remains a black box, as it is impossible to check whether it corresponds to the system described in the original paper. Not knowing the reliability of the system limits its use for distant reading analyses, i.e. quantitative studies that go beyond text exploration.
- **Language Versions.** Currently, TagMe is only available in English, German and Italian but does not support other widespread languages such as Chinese, Arabic, Spanish, and French, which are essential for enhancing its use in the DH community.
- **Infrequent Updates.** TagMe has been initially created on the English 2009 version of Wikipedia and it has been updated only twice (summer 2012, summer 2016). Imagine a setting where a scholar intends to analyze a collection of mainstream news on the Middle East: before the most recent update the system was not able to detect mentions of “Al-Nursa Front”, the former Syrian branch of al-Qaeda.
- **Wikipedia as Knowledge Base.** TagMe, as well as other entity-linking solutions, relies on the assumption that the entries and structure of Wikipedia provide us with a comprehensive and accurate knowledge base. While this is mostly true for standard NLP and IR approaches, when it comes to humanities research this assumption shows all its limitations. As a matter of fact, linking to Wikipedia is not ideal for example when dealing with historical documents, simply because entities and concepts relevant in the corpus may be missing from such a general-purpose knowledge resource (as remarked in Lauscher et al., 2016).

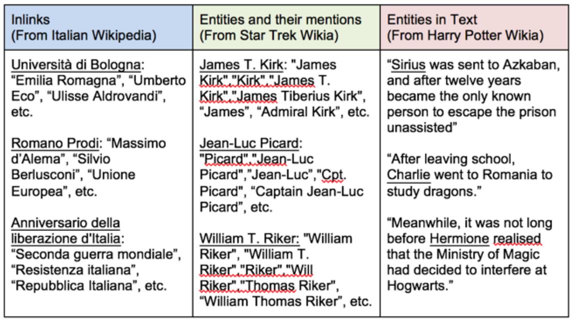
Specific contribution

While we are currently working on the implementation and optimization of a domain-adaptable entity linking pipeline, at the conference we intend to present a solution for generating, in an automatic fashion, domain-specific knowledge bases from an user-created Wiki. As the creation of a complete Wiki is too time-consuming, these domain-specific wikis are used in combination with general world knowledge available on Wikipedia. In particular, we will describe how our system can make use of the following input:

The XML Dump of any language version of Wikipedia and rapidly create the indexes that compose the knowledge base. This permits to have a knowledge base for each language version of Wikipedia and to update it on the spot whenever needed.

Any MediaWiki website dump, such as Wikia (although it is important to consider the copyright license when downloading and using this data), to be merged into the same index. In the table we report a few examples from different Wikia sites. It is important

This solution gives the scholar the possibility of creating (or improving an already existent) domain specific Wikia (a practice common in DH education, see Farabaugh, 2007 and Giglio & Venecek, 2009) on the topic she/he intends to study and identifying mentions of domain-specific and general-purpose concepts in large text collections.



Inlinks (From Italian Wikipedia)	Entities and their mentions (From Star Trek Wikia)	Entities in Text (From Harry Potter Wikia)
Università di Bologna: "Emilia Romagna", "Umberto Eco", "Ulisse Aldrovandi", etc. Romano Prodi: "Massimo d'Alema", "Silvio Berlusconi", "Unione Europea", etc. Anniversario della liberazione d'Italia: "Seconda guerra mondiale", "Resistenza italiana", "Repubblica Italiana", etc.	James T. Kirk: "James Kirk","Kirk","James T. Kirk","James Tiberius Kirk", "James", "Admiral Kirk", etc. Jean-Luc Picard: "Picard","Jean-Luc Picard","Jean-Luc","Cpt. Picard", "Captain Jean-Luc Picard", etc. William T. Riker: "William Riker", "William T. Riker","Riker","Will Riker","Thomas Riker", "William Thomas Riker", etc.	"Sirius was sent to Azkaban, and after twelve years became the only known person to escape the prison unassisted" "After leaving school, Charlie went to Romania to study dragons." "Meanwhile, it was not long before Hermione realised that the Ministry of Magic had decided to interfere at Hogwarts."

Bibliography

Auer, S., et al. (2007) "Dbpedia: A nucleus for a web of open data." *The semantic web*. Springer Berlin Heidelberg, 2007. 722-735.

Cornolti, M., Ferragina, P., and Ciaramita, M. (2013) "A framework for benchmarking entity-annotation systems." *Proceedings of the 22nd international conference on World Wide Web*. ACM.

Farabaugh, R. (2007) "'The isle is full of noises': Using wiki software to establish a discourse community in a Shakespeare classroom." *Language Awareness* 16.1: 41-56.

Ferragina, P., and Scaiella, U. (2010) "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)." *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM.

Giglio, K., and Venecek, J. (2009) "The Radical Historicity of Everything: Exploring Shakespearean Identity with Web 2.0." *Digital Humanities Quarterly* 3.3.

Hasan, K. S., and Ng, V. (2014) "Automatic Keyphrase Extraction: A Survey of the State of the Art." *ACL (1)*.

Hasibi, F., Balog, K, and Bratsberg, S. E. (2016) "On the Reproducibility of the TAGME Entity Linking System." *European Conference on Information Retrieval*. Springer International Publishing, 2016.

Lauscher, A., et al. (2016) "Entities as topic labels: combining entity linking and labeled LDA to improve topic interpretability and evaluability." IJCol-Italian journal of computational linguistics 2.2 (2016): 67-88.

Nadeau, D., and Satoshi S. (2007) "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30.1: 3-26.

Navigli, R., and Ponzetto, S. P. (2012) "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." *Artificial Intelligence* 193 (2012): 217-250.

Text Mining Eighteenth-Century Travel Writing: The "I" and Autobiography of Narration

Catherine Nygren
catherine.nygren@mail.mcgill.ca
McGill University, Canada

How can text mining further explore and challenge our conception of how the style of British travel writing changed from 1700 to 1830? In this poster, drawn from the first chapter of my dissertation, I use text mining methods to explore the purported coming forth of the self in eighteenth-century travel writing, examining texts ranging from instructions for gentlemen and antiquarian accounts to the documents of New World explorers and the journals of female travelers. Increasing public demand for travel-related texts and the genre's potential for experimenting and examining social, philosophical, and aesthetic ideas elevated travel literature to one of the most important - and popular - genres of the century. Despite the modern broadening of the definition of travel literature and analyses that use the lenses of feminism and postcolonialism, however, overviews of travel writing still primarily use close readings and representative samples of texts to make claims about the genre as a whole.

My project, however, will use a corpus of more than 3500 texts (curated through a combination of existing bibliographies and machine learning methods) to examine and challenge our received literary histories of the genre. In particular, this presentation focuses on the "I" and autobiography in narration. In the first half of the eighteenth century, authors crafted travel texts to have a balanced amount of personal narration: the voice of the narrator was required to provide order, entertainment, and authority in the narrative, but with an over-emphatic voice, the author would appear egotistical, fantastical, and unreliable. By the middle of the century, however, scholars identify a trend of authors increasingly writing in a first-person voice and interpolating their detached, factual observations with their personal experiences and reflections. Around the turn of the century, when many locations, particularly Grand Tour destinations like France and Italy, had been described multiple times, readers desired more original writing and critics used personal narration as evidence that the author had actually travelled. Still, authors often felt they had to justify writing in a distinctive voice: Mary Wollstonecraft, when she found she could not "avoid being continually the first person—'the little hero of each tale,'" became "determined to let [her] remarks and reflections flow unrestrained" (3).

In this poster, therefore, I seek to examine this purported coming forth of the self in eighteenth-century travel literature. Tracking distributions of pronouns, such as “I,” “me,” and “you,” and other narrative signifiers will reveal the shifting trends of personal narration in my corpus. Analyzing the words used in titles—such as “Observations,” “Reflections,” and epistolary markers, among others—may provide further insight into how authors and publishers conceptualized and marketed the relationship between the author, the narrator, and the reading public. In examining such patterns, I seek to answer questions such as whether the pieces we consider more “literary” use the first-person voice more often or more experimentally, whether particular sub-genres or texts in my corpus are more likely than others to follow this trend, and whether any other shifts in grammar and vocabulary accompanied this new privileging of first-person narration. A corollary question is whether Sterne’s *Sentimental Journey* is, as several critics contend, truly a catalyst for this trend, and how the influence of Sterne and other significant authors manifest in my corpus. In particular, I will pay close attention to female writers and authors of colour, following the work of critics such as Elizabeth Bohls, in order to see how they subvert and support the trends of travel writing at that time.

This first chapter on the self, along with my dissertation’s examinations of descriptive language and subgenre, will provide an extensive understanding of eighteenth-century British travel writing. My approach encourages both breadth and depth in subject material, from patterns across hundreds of texts to significant paragraphs, sentences and words. This mixture of distant and close reading will offer a reformulation of the massive genre by, for the first time, analyzing a significantly larger corpus than was previously possible. My findings and methodology (shared through Github and other public forums) will provide critical resources and frameworks for future investigations of the language and genre of other forms of eighteenth-century literatures.

Construction of the “Corpus of Historical Japanese: Meiji–Taishō Series I – Magazines”

Toshinobu Ogiso
togiso@ninjal.ac.jp
National Institute for Japanese Language and Linguistics
Japan

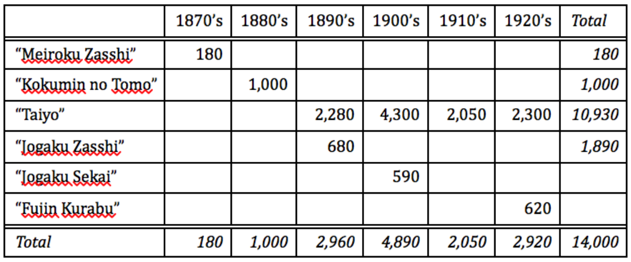
Asuko Kondo
kondo@ninjal.ac.jp
National Institute for Japanese Language and Linguistics
Japan

Yoko Mabuchi
mabuchi@meiji.ac.jp
Meiji University, Japan

Noriko Hattori
nhattori@ninjal.ac.jp
National Institute for Japanese Language and Linguistics
Japan

In this talk, we wish to discuss the construction of the corpus “Meiji-Taishō Series I - Magazines,” (hereinafter called CHJ-Magazines) which was released as a part of the Corpus of Historical Japanese (CHJ). This corpus contains magazines published in Japan in the Meiji and Taishō periods (1868–1911) representative of Modern Japanese language, with the total number of words used in the text reaching a size of some 14,000,000 items. This corpus is the first large-scale corpus with morphological information of Modern Japanese and will contribute to research into that period of the Japanese language.

At the National Institute for Japanese Language and Linguistics, a joint research project entitled “Construction of a Diachronic Corpus and New Developments in Research of the Japanese Language” was begun in 2016. CHJ Magazines is one part of this project, using a number of representative magazines for each of several periods, with the selected published material spread out over set time intervals. The result is a large-scale corpus that makes it possible to examine the state of the written Japanese language in the Meiji and Taishō periods, as well as to examine how the language underwent changes in those periods (Table 1).



	1870’s	1880’s	1890’s	1900’s	1910’s	1920’s	*Total*
“Meiroku Zasshi”	180						*180*
“Kokumin no Tomo”		1,000					*1,000*
“Taiyo”			2,280	4,300	2,050	2,300	*10,930*
“Jogaku Zasshi”			680				*1,890*
“Jogaku Sekai”				590			
“Fujin Kurabu”						620	
Total	*180*	*1,000*	*2,960*	*4,890*	*2,050*	*2,920*	*14,000*

Table 1. Corpus Size for Each Year of Publication (Units: Thousand Words)

One characteristic of this corpus is that morphological (word) information is annotated to each text. In order to annotate highly accurate morphological information to the corpus, it was necessary to develop and utilize a dictionary that could enable automatic morphological analysis for the colloquial and literary styles that was mixed in the written Japanese of the aforementioned periods. We customized UniDic, a dictionary for morphological analysis of Japanese, which can lemmatize variations of orthography and word forms. We also used MeCab, a state-of-the-art Japanese morphological analyzer with this customized UniDic.

While the accuracy of automatic analysis is high at approximately 93%, it is extremely difficult to append morphological information uniformly without any mistakes in such a large amount of text. To address this difficulty, it was necessary to establish a "core" data set, for which we could guarantee the attachment of highly accurate morphological information (with accuracy higher than 99%) through the utilization of automatic computer analysis together with manual (human) correction. For core data, 500,000 words were selected, taking into consideration the balance of publication year, style and genre of each of the articles sampled. Core data is suitable for Japanese research which requires highly accurate morphological information. This data was also used as training data for the dictionary for morphological analysis described above. The automatic analysis results of the "non-core" data set, meanwhile, received a certain degree of manual correction which, while not being exhaustive, strikes a balance between quality and volume.

This corpus is made publicly accessible by way of an online search application called "Chūnagon." With "Chūnagon," it is possible to carry out searches that specify complex combinations of different morphological information (lemma, part-of-speech, conjugation type, lexical strata, etc.). The search results also display information on the source (magazine title, article title, year of publishing), author (name, sex, year of birth), data type (core, non-core), text type (conversation, quotation, other), text style (literary, colloquial, mixture, verse), etc. The author information includes a link to catalogues in the National Diet Library. By making use of this convenient service, it is possible to check how to read the name of the author of the article in question, to check when he or she was born and died, and, at the same time, to see if he or she made use of multiple other names. The "Chūnagon" search results also provide links that allow the viewing of images of corresponding pages in the original magazines.

Bibliography

Center, National Institute for Japanese Language and Linguistics (2014 - 2016) *Corpus of Historical Japanese*. http://pj.ninjal.ac.jp/corpus_center/chj/overview-en.html

Ogiso, T. (2014), "Dictionary and Morphological Analysis in the Historical Corpus." *Nihongogaku*, 432, pp. 83-95. Meiji Shoin. (In Japanese)

Mabuchi, Y., and Asuko, K. (2016), "Explanatory notes of symbols in the text, Display categories for Chunagon: Corpus of Historical Japanese, Meiji – Taisho Series I – Magazines (Short Unit Ver. 0.9)". http://pj.ninjal.ac.jp/corpus_center/chj/doc/abstract-meiji-taisho-2016.pdf (In Japanese)

Cast Your Net Wide: Finding Historical References in Parliamentary Data

Alex Olieman
alex@olieman.net
University of Amsterdam, the Netherlands

Kaspar Beelen
kasparvonbeelen@gmail.com
University of Amsterdam, the Netherlands

Introduction

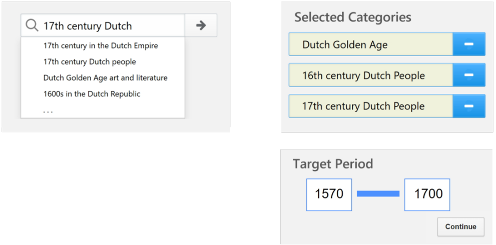
Analyzing how the past is molded to fit contemporary needs and priorities represents an intensely researched topic among historians, who have scrutinized (among others) the performance of "commemorative practices," the transmission of "collective memories," or, in general, "the language of the past." However, scholars currently lack adequate digital tools to assist them in this endeavor (Likaka, 2009; Piersma and Ribbens, 2013; Wilson, 2016).

The information needs of historians and other humanities scholars are not always adequately captured by traditional full-text search because query terms often serve as a proxy for more abstract concepts, while both items (the word and the concept) sometimes hardly overlap. This technology demonstration focuses on a specific type of concept, namely historical eras or periods.

Because of their complex and heterogeneous character, historical periods are difficult to capture with simple keywords. For example, the historian looking for references to the Renaissance in diachronic corpora will often encounter documents in which the term appears in its metaphorical sense, e.g. "the renaissance of train travel," instead of referring to the historical period. Besides yielding many irrelevant hits, keyword search also tends to overlook many relevant documents, because it ignores entities that are related to the query at a *conceptual* level, such as the painter Titian or the Nonsuch Palace.

Traditional full-text retrieval methods will obtain low recall and precision, when the target of a search comprises complex phenomena such as historical events or periods. To overcome these shortcomings, we introduce "WideNet," a tool that allows historians and other scholars within the humanities to search for information about specific periods in a multilingual corpus containing the Parliamentary debates from the United Kingdom, Canada, and the Netherlands.

WideNet models historical periods as a container of (hierarchically) related entities. The art historian, interested in how British MPs refer to the Italian Renaissance will, instead of retrieving all occurrences of the term “Renaissance,” be provided with speeches that mention, e.g., composers such as Costanzo Porta, and famous paintings such as The Bacchanal of the Andrians. Whereas earlier work proposed to use such entities as search suggestions (Piersma et al., 2014), we rather prefer a “sculptor’s approach” in which many containers of potentially relevant entities are initially included in the query, and may be deselected based on their empirical relevance.

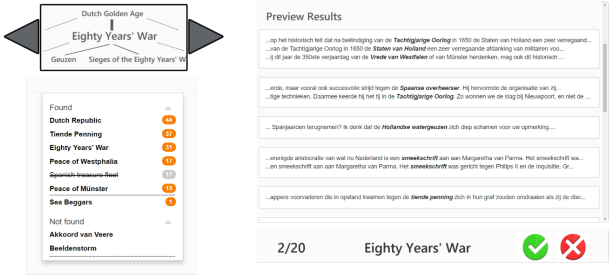

17th century Dutch
17th century in the Dutch Empire
17th century Dutch people
Dutch Golden Age art and literature
1600s in the Dutch Republic
. . .

Selected Categories
Dutch Golden Age
16th century Dutch People
17th century Dutch People

Target Period
1570 1700

Figure 1: Initial query specification in WideNet

These categories of entities are provided by a knowledge base (KB) that is loaded into the tool in advance. Our demo tool makes use of DBpedia, but any KB that conforms to the SKOS ontology may be loaded. The scholar, using WideNet, starts by selecting one or several root categories from a type-ahead search box (see Figure 1), and can further demarcate the query by selecting a time period, which will be used to prune the underlying entities of the selected categories. WideNet subsequently retrieves the network of narrower categories for each selected root category, and collects the contained entities as potentially relevant query components.



Found
Dutch Republic 44
Tiende Penning 37
Eighty Years' War 31
Peace of Westphalia 17
~~Spanish treasure fleet~~ 17
Peace of Münster 15
Sea Beggars 1

Not found
Akkoord van Veere
Beeldenstorm

Preview Results

...op het historisch feit dat na beëindiging van de ***Tachtigjarige Oorlog*** in 1650 de Staten van Holland een zeer verregaand...
...van de Tachtigjarige Oorlog in 1650 de ***Staten van Holland*** een zeer verregaande afdanking van militairen voo...
...ij dit jaar de 350ste verjaardag van de ***Vrede van Westfalen*** of van Münster herdenken, mag ook dit historisch ...

...erde, maar vooral ook succesvolle strijd tegen de ***Spaanse overheerser***. Hij hervormde de organisatie van zij...
...tige technieken. Daarmee keerde hij het tij in de ***Tachtigjarige Oorlog***. Zo wonnen we de slag bij Nieuwpoort, en niet de ...

... Spanjaarden terugnemen? Ik denk dat de ***Hollandse watergeuzen*** zich diep schamen voor uw opmerking....

...erenigde aristocratie van wat nu Nederland is een ***smeekschrift*** aan aan Margaretha van Parma. Het smeekschrift wa...
...en smeekschrift aan aan Margaretha van Parma. Het ***smeekschrift*** was gericht tegen Philips II en de Inquisitie. Gr...

...appere voorvaderen die in opstand kwamen tegen de ***tiende penning*** zich in hun graf zouden omdraaien als zij de disc...

2/20 Eighty Years' War ✓ ✗

Figure 2: Assessing the relevance of categories and entities

The next step for the WideNet user is to assess which of the retrieved subcategories actually contain entities that lead to relevant results. The interface facilitates this task by showing, per subcategory, which entities are mentioned in the corpus, and how frequently, as well as which entities did not occur (see Figure 2). It also displays a list of preview results, showing limited context, to offer quick clues about the relevance of the category. This preview is also useful to identify individual entities that are not relevant after all, which can be deselected.

After inspecting and selecting relevant categories of entities, the demo interface allows further exploration by providing an environment in which the retrieved documents can be tagged and subjected to close reading. Moreover, the user can examine the results in relation to the parliamentary metadata, i.e. look for saliency by plotting the annotations over time, or study bias by comparing how often different political parties refer to the entities of interest.

WideNet offers a flexible and widely deployable interface that enables scholars to simultaneously search for myriad aspects that have shaped specific historical eras. It provides researchers with a holistic picture of how these periods are discussed in parliament, and thereby helps future scholars to get a more profound understanding of how history ties in with contemporary issues, and how societies deal with their past.

Bibliography

Likaka, O. (2009). Naming Colonialism: History and Collective Memory in the Congo, 1870–1960. Madison: University of Wisconsin Press.

Piersma, H., Tames, I., Buitinck, L., Doornik, J., and Marx, M. (2014). “War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History.” Digital Humanities Quarterly, 8(1).

Piersma, H. and Ribbens, K. (2013). “Digital Historical Research: Context, Concepts and the Need for Reflection.” BMGN - Low Countries Historical Review, 128(4): 78–102. DOI: http://doi.org/10.18352/bmgn-lchr.9352

Wilson, R. (2016). The Language of the Past. London: Bloomsbury Academic.

Textograf: A Web Application for Manuscript Digitization

Boris Orekhov
nevmenandr@gmail.com
National Research University
Higher School of Economics, Russia

Fekla Tolstoy
6975991@gmail.com
The Leo Tolstoy State Museum, Russia

Textograf is a web-based app for the digitization of manuscripts. *Textograf* is intended to automate the work of the textual critic of any handwritten text. *Textograf* allows you to select a specific set (or “layer”) of edits to a text, to compare different iterations of the text, and ultimately to visualize the correlation between the final text and the process that stood behind it. *Textograf* is able to work with prose and poetry, with manuscripts and printed sources. A description of the project can be found on the *Textograf* web page.

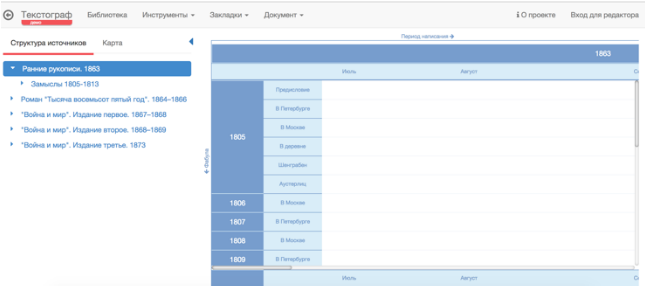
The app allows you to upload an image of the manuscript, to enter the transcript of the manuscript, and to match specific parts of the transcript with where they appear on the page of the manuscript. You can also mark on the transcript the different stages of edits that were made. The app also enables you to see and highlight all text that was deleted, all text that was added, and all text that was initially deleted and subsequently restored.

All documents (scanned images and texts) are stored in a so-called “library”. Access to the documents depends on the documents’ status – “public” or “private”. Private documents are visible only to the editor, public documents can be accessed by everyone.

Textograf allows its user to create complex documents that include manuscript images and their text versions. Text fields are searchable, allowing the user to find search results in the manuscript itself. Every manuscript image can be archived according to customized categories (whose handwriting, where it is currently stored, size of paper, pen or pencil, published or not published, etc).

Textograf allows you to download documents from the library in TEI format. Texts can then be ordered in accordance with the user’s version of the text. Often the text has a number of different versions, and the user has to find the right one and to collect all the corresponding manuscripts. *Textograf* makes this possible.

Different types of source documents connected to the work (outlines, synopses, the manuscripts themselves, printed editions) can be shown as an infographic, demonstrating visually the correlation between the different source documents. For example, this is a map of the early stages of Leo Tolstoy’s work on *War and Peace* displayed on the Textograf site.



The app was initially developed using the manuscripts of Leo Tolstoy: 200 pages (out of 5,000) from *War and Peace*.

The map has two axes: the real time of Tolstoy’s work (starting from 1863) and the fictional place and contents of the novel so you can see which episodes Tolstoy started with and how each part of the text came together.

As of today, the *Textograf* app is unique. It allows the user to work with manuscripts and texts in automatic or semi-automatic mode, meaning that the user can focus on the creative aspects of this work and become immersed in the manuscripts. Before the creation of *Textograf*, many of these functions had to be performed manually, such as establishing the link between text and image, and identifying the sequence of writing.

In the process of developing the application, we had to convert the terms and methods of textual criticism in terms of electronic resources. For example, we had to establish the meaning of the word "layer", or "edit", and to show how to visualize it on the screen.

When we talk about the influence different inputs, we mean that any text has a complex history. For example, a poem may obtain its final form as a result of a long history of changes from one version to another. Thus, what stays in the archives will be several manuscripts that contain similar, but not identical versions of the same text. Earlier versions will influence the later ones. The "text map" shows the full history and evolution of the text.

Textograf is designed to work with any paper-based sources. It could be both manuscripts and printed editions. *Textograf* is a computer programme editor and is designed to prepare in electronic form both critical and diplomatic editions.

Textograf can work with the manuscripts of any writers, be it Leo Tolstoy, William Faulkner, Oscar Wilde, Marcel Proust or anyone else.

As already mentioned, *Textograf* is a web-based application and cannot be installed on a computer. Integration with other tools is possible only through the files that can be downloaded from the application.

A Handbook of Electronic Literature Reading

Élika Ortega
e.ortegaguzman@northeastern.edu
Northeastern University, United States of America

Erik Radio
radio@email.arizona.edu
University of Arizona, United States of America

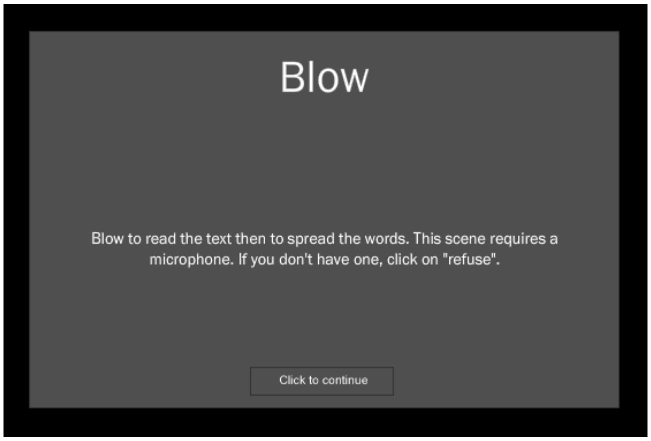
The changes effected by digital platforms on reading continue to be a contested area in popular culture, pedagogy, and higher education. This is also an area where electronic literature (e-lit), digital humanities scholarship, and information science can shed much light. Works of e-lit are a rich testing ground to explore how creative, non-commercial uses of digital media have been exploited and proposed to readers. In our *Handbook of Electronic Literature Reading*, we explore the reading instructions accompanying e-lit works to investigate the performative engagement of reading from the perspective of storage medium, objects and modes of interaction, and gestures or actions. The data included in the reading instructions shed light on how reading practices have morphed depending on platform historically over the last three decades. This focus has warranted an examination into current bibliographic practices to determine how the new modalities of e-lit require different descriptive methods and approaches to metadata creation.

Background

To say that digital reading platforms (web, e-readers, smartphones, etc.) are changing the way reading has been practiced in the age of print is now a truism. Often, these shifts have been the object of elegiac meditations (Birkerts, Manguel) and pedagogical concerns (NEA's "Reading at Risk" and "To Read or Not to Read"). These approaches highlight the 'erosion' that the practice of reading is undergoing. A useful corrective has been Andrew Piper's Book Was There. His call to think about reading from the basis of "the relationship between reading and hands, [and] the long history of how touch has shaped reading and, by extension, our sense of ourselves while we read" (3) has brought to light the bodily handling of reading objects. Our understanding of reading and text as a performative kind of practice has begun to take shape. For Rita Raley, text is "the whole of the event, its physical, logical, and conceptual architecture; the enactment and experience; its temporal structures; and associated social and juridical protocols" (2013: 21). Similarly, for Johanna Drucker, we need to elaborate "a different conception of artifacts (books, documents, works of textual or graphic art), one in which reception is production and therefore all materiality is subject to performative engagement within varied, and specific, conditions of encounter" (2014).

Electronic Literature Instructions

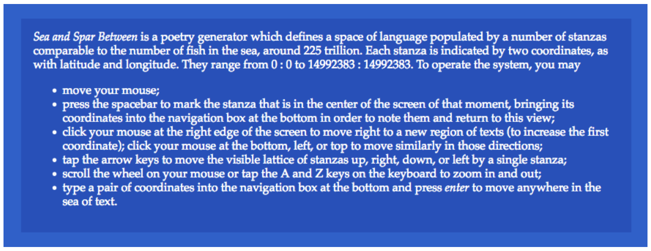
Though a sizeable body of e-lit scholarship has focused on matters like textuality, software, meaning making systems, code, preservation, etc. reading has not been at the forefront. The experimental qualities of e-lit have produced a great diversity of interaction modalities and ways of handling the objects that supports said experimentation. Nevertheless, an understudied convention of e-lit are the instructions accompanying most works. These instructions go from some of the most radically sensuous like Serge Bouchardon's *Blow*, "Blow to read the text then to spread the words. This scene requires a microphone" (Fig 1); to more conventional ones like "click your mouse at the right edge of the screen to move right to a new region of texts [and] tap the arrow keys to move" from Nick Montfort and Stephanie Strickland's *Sea and Spar Between* (Fig 2).



Blow

Blow to read the text then to spread the words. This scene requires a microphone. If you don't have one, click on "refuse".

Fig 1. Screenshot from Serge Bouchardon's Blow



Sea and Spar Between is a poetry generator which defines a space of language populated by a number of stanzas comparable to the number of fish in the sea, around 225 trillion. Each stanza is indicated by two coordinates, as with latitude and longitude. They range from 0 : 0 to 14992383 : 14992383. To operate the system, you may

- move your mouse;
- press the spacebar to mark the stanza that is in the center of the screen of that moment, bringing its coordinates into the navigation box at the bottom in order to note them and return to this view;
- click your mouse at the right edge of the screen to move right to a new region of texts (to increase the first coordinate); click your mouse at the bottom, left, or top to move similarly in those directions;
- tap the arrow keys to move the visible lattice of stanzas up, right, down, or left by a single stanza;
- scroll the wheel on your mouse or tap the A and Z keys on the keyboard to zoom in and out;
- type a pair of coordinates into the navigation box at the bottom and press *enter* to move anywhere in the sea of text.

Fig 2. Screenshot from Nick Montfort and Stephanie Strickland's Sea and Spar Between

The paratextual dimension of these instructions has been on the margins of the studies focusing on particular works. Cumulatively, however, the information kept on these reading instructions can signal the shifts that e-lit has enacted on reading in digital platforms. Even more broadly, as a sample of the many experimental and still unstable standards of reading in digital platforms, e-lit reading instructions can offer evidence of practices that may be becoming more integral to the act of reading. Similarly, the mentions of the hardware (mouse, microphone, keyboard, etc.) required to read e-lit works provide detailed insights on the historical role that technological developments have enacted on reading as a performative engagement.

Metadata and Description

The new affordances offered by e-lit increase the scope of bibliographic practices in ways that allow for more coordinated retrieval but also create greater depth and specificity in records as datasets. Many descriptive metadata standards have been largely shaped by the predominance of print media, and while its qualities overlap with those of e-lit, emerging characteristics of the latter require extensible structures to be considered comprehensively described. E-lit specific metadata sets have been developed by initiatives like Electronic Literature as a Model of Creativity and Innovation in Practice (ELMCIP) and the Electronic Literature Directory (ELD). However, among them there is still a variety of descriptive approaches and objectives. The Consortium on Electronic Literature (CELL) has taken significant steps towards interoperability and a "consensual model for the object of this field" (Baldwin) while still acknowledging the particular interest of individual archives and datasets. Efforts like this may still be enhanced through greater reconciliation with existing bibliographic procedures.

In e-lit the interaction between a work and a reader poses new ontological considerations for descriptive orientations suitable to examine the particularities of reading in digital platforms. Hayles' concept of intermediation proposes that the emergent processes that occur when reading e-lit forms a dynamic heterarchy in which both reader and work continuously inform and shape the trajectory by which reading unfolds (2007:100). This process reframes the ontic nature of work and reader, dissolving the duality into a single event. New qualities surface as a result, highlighting the different modalities through which intermediation occurs and inform bibliographic projects. Navigating these qualities and the language with which they are described poses new challenges for interoperability with similar collections in addition to being synthesized into the extensive legacies of bibliographic practices.

The Handbook

A Handbook of E-Lit Reading collects screenshots of the reading instructions of e-lit works and documents them according to storage medium (web, iPad, CD-ROM, floppy disk, etc.), actions or gestures (blow, click, scroll, type, etc.) and hardware (web cam, keyboard, mouse, microphone, etc.). The pilot target corpus includes ~200 works from over twenty countries anthologized in the Electronic Literature Collections Vol. 1, 2, and 3. The data extracted from the instruction pages will be analyzed and cross-referenced in order to observe emerging patterns like most common practices, most resilient ones, the rise and fall of some of them, as well as the introduction of new technological developments in hardware or software that have marked important shifts in the creation and reading of e-lit. This corpus, though not exhaustive, provides a starting point to design a suitable metadata set and test the categories included in it. Ultimately, the corpus will also encompass enough data to start drawing a hypothesis and identify future directions for the project.

Bibliography

Baldwin, S. (2015). *CELL: On the Consortium for Electronic Literature.* The Center for Literary Computing. http://cellproject.net/documentation

Birkerts, S. (2006). *The Gutenberg Elegies: The Fate of Reading in an Electronic Age.* New York: Faber and Faber.

Boluk, S., Flores, L., Garbe, J., and Salter, A. (2016) (eds). *The Electronic Literature Collection.* Vol. 3. Cambridge Mass: Electronic Literature Organization.

Borràs, L., Memmott, T., Raley, R., and Stefans, B. (2011) (eds). *The Electronic Literature Collection*. Vol. 2. Cambridge Mass: Electronic Literature Organization.

Drucker, J. (2014) "Distributed and Conditional Documents: Conceptualizing Bibliographical Alterities." *MATLIT: Materialidades da Literatura* 2.1: 11–29.

Hayles, N. K. (2007). "Intermediation: The Pursuit of a Vision." *New Literary History* 38.1: 99-125

Hayles, N. K., Montfort, N., Rettberg, S., and Strickland, S. (eds). (2006) *The Electronic Literature Collection.* Vol. 1. College Park, Maryland: Electronic Literature Organization, 2006.

Manguel, A. (1996). *A History of Reading*. London: HarperCollins, 1996.

National Endowment for the Arts. (2004) "Reading at Risk. A Survey of Literary Reading in America." *Research Division Report* No.46. Web. 17 Jan 2012.

National Endowment for the Arts. (2007). "To Read or Not to Read: A Question of National Consequence." *Research Division Report* No.47. : Web. 17 Jan 2012.

Piper, A. (2012). *Book Was There: Reading in Electronic Times.* Chicago: University of Chicago Press.

Raley, R. (2013). "TXTual Practice." Hayles, Katherine, and Jessica Pressman, eds. *Comparative Textual Media Transforming the Humanities in the Postprint Era.* Minneapolis: University of Minnesota Press.

Rettberg, S., and Rasmussen, E. D. (2014). "The ELMCIP Knowledge Base." in *ELMCIP: Electronic Literature as a Model of Creativity and Innovation in Practice.* Morgantown, VA: The Center for Literary Computing.

Death by Numbers: Bills of Mortality in Early Modern London

Jessica Otis
jotis@andrew.*cmu*.edu
Carnegie Mellon University

The foundation of *Death by Numbers: Bills of Mortality in Early Modern London* is the construction of a relational database containing plague mortality information from the London Bills of Mortality in all their varying levels of granularity. Plague first entered England in 1348, as part of a continent-wide epidemic that killed approximately one third of the population of Europe. Thereafter, England suffered continuous outbreaks of plague through 1679, and fear of plague lasted well into the eighteenth century. The most well-documented epidemics of the early modern era were in England's cities, particularly London, which suffered six major epidemics in the century between 1563 and 1665, and lost an estimated 225,000 people to plague. During the mid-sixteenth century, in response to these epidemics, some city officials began to compile numerical summaries of local deaths and circulated those mortality statistics in manuscript form. By the turn of the seventeenth century, there was enough popular interest in these numbers that officials in the city of London began publishing weekly mortality statistics in a series known as the Bills of Mortality. The weekly bills included a parish-by-parish list of total deaths and plague deaths, along with running tallies of city-wide christenings, deaths broken down by a variety of causes, and parishes infected with plague. London's weekly bills were also supplemented annually with a general account of the preceding year, published on the Thursday before Christmas. Over the first half of the seventeenth century, more parishes were added to the bills, including the parishes in surrounding counties and in the neighboring city of Westminster, until the "Bills of Mortality" formed a recognized geographical unit that included all of London and its suburbs.

The Bills of Mortality are thus a vital source of information about the population of seventeenth- and eighteenth-century London. They have been analyzed by politicians, demographers, historians, and epidemiologists as early as the seventeenth century itself, but currently only the bills from a few select epidemic years are available online. Construction of the Bills of Mortality database is thus a vital step for both this project and for future digital humanists interested in the bills. After the construction of this database, this project will use network analysis to examine the transmission of plague between residents of London's parishes during both famous epidemic outbreaks and the periods of endemic plague that followed them. Subjects of particular interest in this analysis include the change in infection patterns over time, how the infrastructural and geographical features of the city act as barriers or "highways" to infection, and the devastating impact of – and slow recovery from – the Great Fire of London while plague was still active in the city. The project will also include a mapping component to visualize mortality trends by parish over time, with particular interest in the full duration of epidemics as opposed to each epidemic's highest mortality (and highest profile) year. Eventually the Bills of Mortality database, as well as materials associated with the network analysis and visualizations, will be made freely available online in accessible, sustainable formats for reuse by other members of the scholarly community.

Bretez II : La machine à remonter le temps

Mylène Pardoen
mylene.pardoen@wanadoo.fr
Institut des Sciences de l'Homme de Lyon, CNRS, France

Le poster met en lumière les collaborations nécessaires pour l'élaboration d'un projet transdisciplinaire complexe portant sur l'élaboration d'une maquette de restitution urbaine en 5D (comprenant, en plus du tridimensionnel, le déplacement et le sensible – ici le sonore) au contenu scientifiquement valide. Il en présente les articulations et les différentes phases de restitution ainsi que les productions possibles suivant les types de supports.

Bretez II est un projet restitution visuelle et sonore de Paris au XVIIIe siècle. Ce projet fait suite à *Bretez*, avec des équipes renouvelées afin de prendre en compte les évolutions et les nouvelles problématiques apparues en cours d'élaboration de la phase initiale. Ce programme de recherche propose un nouvel outil pour des applications en médiation scientifique et culturelle. Entrant dans le cadre des **humanités numériques**, *Bretez II* est projet transdisciplinaire fort de multiples interactions entre SHS (Sciences Humaines et Sociales) et sciences et technologies de l'information et de la communication (STIC – Sciences de l'Ingénieur), qui tous participent à l'élaboration de la maquette virtuelle.

Placé sous le parrainage de Daniel Roche (Professeur Honoraire au Collège de France – Paris), labellisé IMU (Intelligence des Mondes Urbains – LabEx de la région Rhône-Alpes – France), accompagné par la SATT (société d'accélération de transfert technologiques) PULSALYS (organisme de valorisation sur le bassin lyonnais), *Bretez II* est riche de ses nombreux partenariats. Outre les partenaires issus de la Recherche, nous pouvons également citer : le Musée Carnavalet (Paris-France), les Archives Nationales (Paris-France), pour les organisations institutionnelles, Labo-M, société franco-berlinoise de production de produits transmédia innovants et la start-up ASPIC (Tourcoing – France).

La zone restituée est un quartier de Paris aujourd'hui totalement disparu – 25 hectares situés entre l'apport de Paris, le Pont au Change, la rue de la Pelleterie et le quai de Gesvres, soit près la Place du Châtelet actuelle.

Ce projet se dénote par une **nouvelle approche** de la restitution du passé en **5D** (la combinaison du visuel [la 3D habituelle], le déplacement à la première personne [FPS] et la dimension sonore – le tout permettant une immersion, donc un meilleur ressenti) en s'appuyant sur une trame sonore de type hétérographique. La dimension **sonore**, cœur du projet, permet de rendre le passé disponible et tangible pour un très large public – complétant une vision méconne et scientifiquement valide de Paris. Il prend appuie sur une plateforme de jeu (Unity 3D et un middleware (Wwise).

Sa finalité est d'élaborer un **nouveau modèle** de restitution historique à destination large, dont la muséographie et la recherche : une matrice qui permette de se soustraire à l'obsolescence des supports et de permettre une diffusion sans renouvellement d'installations souvent coûteuses l'investissement. Initialement à destination des musées et des sites patrimoniaux, Bretez cible d'autres applications : **produits ludo-éducatifs** ou culturels sur supports nomades ou non, livres numériques, réalité virtuelle et augmentée, *in situ* ou non.

Ses spécificités :

- Une approche originale ;
- Le sensible comme vecteur complémentaire de recontextualisation ;

- Une dimension sonore cohérente avec l'histoire sociale et urbaine du quartier et la géométrie des architectures modélisées rendant sa prise en compte prééminente.

Les innovations du concept :

- La 5D : la 3D, le déplacement à la première personne [FSP] et la prise en compte de la dimension sensible – le sonore ;
- La modélisation volumique riche en détails – tant pour les intérieurs que pour les extérieurs des bâtiments – et leur génération en temps réels.

Les innovations concernant l'édition scientifique :

- Des nouveaux procédés éditoriaux de type hétérographiques ;
- Création d'un nouveau média impliquant la réalité augmentée.

Genealogies of VR in the archives

Elizabeth Parke
elizabeth.parke@mail.utoronto.ca.
University of Toronto, Canada

Liz Ridolfo
liz.ridolfo@utoronto.ca
University of Toronto, Canada

Using stereoscope cards and a stereoscope viewer from 1900, held in the Thomas Fisher Rare Book Library and the University of Toronto Mississauga Library, this poster presents the history of early immersive viewing experiences by tracing the history of 3D ways of seeing from the stereoscope cards to Oculus Rift and Google cardboard. Secondly, we offer tools for teaching with archival photographic materials, examining the histories of seeing, and integrating digital tools in the curriculum to enrich students' access to these cards' histories while also introducing recent VR development tools such as Unity, and Cardboard. Lastly, by using 3D scanning and printing we replicate the viewer, further expanding possible student engagement with DH tools, methods, and theories. The resulting student projects can be archived in a virtual exhibit (using Omeka), and the meta-data and data from the cards and viewer added to the objects' records.

The goal of this pilot project is to develop an 'out of the box' teaching resource for use in the classroom that offers analogue and digital engagement with rare materials, questions the histories of seeing stereoscopically, and contextualizes contemporary devises for seeing this way in a longer history of photography and immersive media that is often overlooked. Our project pushes students and researchers to consider the longer genealogy of virtual reality viewers and draws on previous work such as that of the Stereogranimator at the New York Public Library lab. In so doing, we seek to recover early histories of immersive media, photography, printing, and popular entertainment in the early 20th century.

Understanding Narrative: Computational Approaches to Detecting Narrative Frames

Eva Portelance
eva.portelance@mail.mcgill.ca
McGill University, Canada

Andrew Piper
andrew.piper@mcgill.ca
McGill University, Canada

Understanding narrative structure at a large scale remains a challenging problem within the field of cultural analytics and computational linguistics. Our aim with this project is to develop novel methods to study the pacing of narrative scene changes and the overall distribution of different plotlines within novels. Being able to analyze such narrative features at large scale can give us insights into the way different genres, time periods, or cultures favor different modes of storytelling. In this project we formalize definitions of narrative scenes and implement new methods of detection and clustering using computational methods.

The project involves three steps: creating an operational concept of a narrative frame; the algorithmic segmentation of narratives by frames; and the predictive clustering of frames into larger-scale "plotlines."

We define a "frame" as a significant shift of three variables in a given text window: entities, actions, and objects, which we represent using POS tagging as proper names, verbs, and nouns. We measure significant lexical shift of our three primary variables in a sliding textual window of 1000 words with increments of 100-word shifts. We test window-size and variable selection relative to human annotation to determine the best performing model. We resolve frames into "plotlines" using hierarchical clustering, also demonstrated in our poster. "Frames" serve as inferred textual units and "plotlines" as aggregated clusters of frames.

We have tested the performance of different combination of variable selection on nine 12,000-word passages from novels of different genres from a range of time periods (from 1818 to 2011). To date, our algorithm outperforms the current state-of-the-art in Hearst's Texttiling algorithm

(1994; 1997) when it comes to placing breaks in the narrative event progression. The performance of our system relative to human performance on the same task (F1 82%, Precision 81%, Recall 86%), shows an F1 score of 69% with 71% precision and 67% recall, where ⅔ annotators agreement is considered to be a true frame boundary. Hearst's method applied to the same problem performs at a significantly lower rate (F1 18%, Precision 18%, Recall 19%). While the overall problem remains challenging we show significant improvement over state of the art systems at detecting narrative segments. Nevertheless, the imperfect accuracy suggests that scene changes have a number of subtle variables that are not exclusively tied to vocabulary or character shifts, which indicate further avenues for future research.

Our poster will present our formalization of narrative events, the results and approach of the segmentation task and the clustering models used. We see this project as a crucial contribution to the larger study of narrative form across different literary genres and time periods.

Bibliography

Hearst, M. (1994). "Multi-Paragraph Segmentation of Expository Text." *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, Los Cruces, NM.

Hearst, M. (1997). "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages." *Computational Linguistics* , 23 (1), pp. 33-64.

VisColl: Modeling and visualizing the physical structure of medieval manuscripts, a poster and demonstration

Dot Porter
dot.porter@gmail.com
University of Pennsylvania Libraries
United States of America

Alexandra Gillespie
alexandra.gillespie@utoronto.ca
University of Toronto, Canada

Alberto Campagnolo
alberto.campagnolo@gmail.com
Library of Congress, United States of America

Laura Mitchell
laura.mitchell@utoronto.ca
University of Toronto, Canada

Rachel Di Cresce
rachel.dicresce@utoronto.ca
University of Toronto, Canada

VisColl is a data model and associated tools that are designed to help scholars to visualize the physical collation of medieval manuscripts. In manuscript descriptions and library catalogs, a collation is normally given in the form of a formula, which describes each quire in terms of the position of that quire in the manuscript, how many leaves the quire contains, and if any leaves have been added or removed. A diagram may also be used to illustrate the same information, with the added benefit of clearly showing which leaves are conjoined (conjoined leaves are also known as bifolia). VisColl enables scholars to model the collation of manuscripts and then to present that information in various ways, including diagrams and formulas, but also in novel ways distinct from collating a manuscript by hand. For instance, in addition to visualizing the physical structure of a manuscript, the Beta Version of the VisColl data model currently under development enables users to create taxonomies describing the content of the manuscript, and other elements, which will enable tools to link those taxonomies to the physical structure, producing a more robust and descriptive visualization than is possible in the current system.

VisColl was conceived in the mid-2000s by Dot Porter during her work at the Collaboratory for Research in Computing for Humanities at the University of Kentucky (UKY). Porter developed the tool in order to address issues she encountered in effectively visualizing standard descriptions of manuscripts in scholarly works. For instance, in *Beowulf and the Beowulf Manuscript* Kevin Kiernan uses the physical construction of the manuscript to make arguments about the dating of the text (separate from the dating of the manuscript itself). In addition, Ben Withers (of UKY), in *The Illustrated Old English Hexateuch, Cotton MS. Claudius B.IV: the Frontier of Seeing and Reading in Anglo-Saxon England*, similarly used a detailed collation statement of the manuscript as the backbone for his investigation of the construction of the manuscript. There are numerous examples of scholarly works that build an argument about the dating and construction of manuscripts based on the collation of the physical object. In consulting such works, Porter saw an opportunity to enable readers to better visualize the structure of the object beyond the limitations of traditional formulas, diagrams, and collation statements.

Digitized medieval manuscripts are typically viewed through page-turning interfaces, which give the impression of page openings, but lack the physical cues present in a physical book, i.e., the size of the book, its thickness, details of the parchment or paper, etc. Indeed, page-turning interfaces do not usually show a picture of book openings at all, but rather they are composites made with two images: one of the left-side page and another of the right-side page. These images would have been taken at different times. Typically all images of one side pages are taken first, e.g. all the rectos, then of the other side, and then file names or structural metadata are used to order the files correctly in post processing. Most digital libraries provide some information on the pages depicted, and views other than single pages: all provide information on the folio number and the side (recto or verso) shown; some indicate the quire number, and some offer a variety of viewing modes, such as single pages, double pages, pages of thumbnails or thumbnails presented filmstrip-style across the bottom of a page. However, again, for the most part, the focus of these resources is on the page, rather than on the physical object. Even the Turning the Pages™ software, conceived by the British Library in 1996 (and developed by Amarillo Systems since 2001), which, since version 2.0 (2006), has produced realistic three-dimensional books (including the ability to mimic the different movement of paper and parchment pages as these are turned), lacks any modelling of the gathering structure. To present knowledge, there is no institutional digital library that describes the physicality of manuscripts outside of the standard Physical Description section of the manuscript records and collation formulas.

The Alpha Version of the VisColl data model is implemented in the Collation Modeler and the Collation Visualizer hosted at the University of Pennsylvania, but the data model was envisioned as agnostic and was designed to be easily used by other collation tools. The University of Toronto, through a Mellon-funded project entitled Digital Tools for Manuscript Study, is developing a robust VisColl web application which implements the Beta Version of the data model, and allows users to visually manipulate and present diagrams and metadata in real time, while also making use of the International Image Interoperability Framework (IIIF) to integrate digital manuscript images alongside scholarly work.

This poster will document the stages of the development of VisColl, from its conception to its current instantiation, highlighting the steps taken and the reasoning behind each new actualization of the project, and will also serve as a demonstration of the current version of the tool developed by the University of Toronto Libraries. Documentation and code of this version can be found and downloaded from the University of Toronto Libraries' GitHub page. The current state of development can be found at VisColl's GitHub page, which documents each new build, and from which the project's code can be downloaded.

Bibliography

Kiernan, K. S., and Prescott, A. (1996). *Beowulf and the Beowulf Manuscript*. London: British Library, Print.

Withers, B. C. (2007) *The Illustrated Old English Hexateuch, Cotton Claudius B.iv: The Frontier of Seeing and Reading in Anglo-Saxon England*. London: British Library. Print

Armadillo Systems (n.d) Turning the Pages™. "Turning the Pages." Web. 01 Nov. 2016. <http://ttp.onlineculture.co.uk/>

Porter, D. (2016). VisColl. University of Pennsylvania, 23 Oct. 2016. Web. 07 Apr. 2017. <https://github.com/leoba/VisColl>

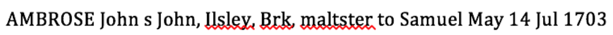
University of Toronto Libraries. (n.d.) University of Toronto. <https://github.com/utlib>

The London Brewers' Prosopography: Parsing Historical Apprenticeship Records

Harvey Quamen
hquamen@ualberta.ca
University of Alberta, Canada

This poster describes one preliminary aspect of a new project about the history of beer and brewing in London. I am building a prosopography of brewers' apprentices in the years from about 1530 to approximately 1800. A catalogue of apprentices and their masters can teach us not only about the social and cultural history of British beer during that time but about the social network of people working in the industry. During this 270-year period, the Worshipful Company of Brewers, the medieval brewing guild first established by Henry VI in 1438, logged nearly 10,000 apprenticeship records (Webb), a collection that serves as just one of many potential datasets that can yield insights into England's brewing culture.

The more immediate goal described in this poster, then, is how to parse these 10,000 records into component parts— people, places, occupations, and dates—so that these relationships can be analyzed and mapped over time. A typical apprenticeship record looks something like this:



AMBROSE John s John, Ilsley, Brk, maltster to Samuel May 14 Jul 1703

In other words: John Ambrose, whose father is John Ambrose from Ilsley, Berkshire and is a maltster by profession, was apprenticed to Samuel May and the fee was paid on 14 July 1703. The record lists three people, a parish, a county, a profession, and a date—a typical dataset found when tracking apprenticeships (Lane).

The simplicity and regularity of that template tantalizingly suggests writing an automated parser, which I am doing with an open source Python module called *pyparsing*. Although these recursive descent parsers, as they are called, are designed for more elaborate projects (like writing compilers), they are the perfect tool for a job like this because they allow users to construct grammatical "rules" that look like simple additions. For example, the record above can be parsed according a grammatical rule that looks like simple Python:

```
apprentice + father + location + occupation + to + master + date
```

But the wide variety of apprenticeship records presents a challenge. As it turns out, the first 10% of the apprenticeship records require nearly 40 different template “grammars.” The effectiveness of an automated parsing approach—useful but nonetheless somewhat limited—is the main point of the poster. Supplementary strategies (like natural entity parsing and dictionary lookups) may provide some help and, if they prove themselves worthy, they will become part of the presentation as well.

Bibliography

Lane, J. (1996). *Apprenticeship in England, 1600-1914.* London: University College London Press.

Pyparsing. http://pyparsing.wikispaces.com/.

Webb, C. (1996). *London Apprentices: Brewers’ Company, 1685-1800.* Volume I. London: Society of Genealogists.

Webb, C. (2001). *London Livery Company Apprenticeship Registers. Brewers’ Company, 1531-1685.* Volume 36. London: Society of Genealogists.

The Use of Cognitive Digital Games in School: Contributions to Attention

Daniela Karine Ramos
dadaniela@gmail.com
Universidade Federal de Santa Catarina, Brazil

Bruna Anastacio
brunaanastacio@hotmail.com
Universidade Federal de Santa Catarina, Brazil

Cognitive games involve a number of different games working aspects of human cognition, while proposing the intersection between the sets of concepts, fun and cognition, for the improvement of cognitive functions. The attention is the main point made in this study, since it is fundamental to the learning process and be recurring complaint among parents and teachers in schools.

With respect to the contributions of digital games to improvement of cognitive processes, researchers suggest that regular practice has a significant influence on improving the performance related to basic visual skills (Li, Polat, Scalzo, & Bavelier, 2010); on the ability to perceive objects simultaneously (Dye & Bavelier, 2010; Feng, Spence, & Pratt, 2007); and on the ability to do more than one task at the same time (Boot, Kramer, Simons, Fabiani, & Gratton, 2008). Other studies specifically investigate the use of digital games in the school context and suggest potential for digital game use to improve of student's attention span at preschool age (Rueda, Checa, & Cómbita, 2012), to improve overall intelligence capacity of elementary school children (Miller & Robertson, 2010), and to better performance of working memory ability (Klingberg et al., 2005; Thorell, Lindqvist, Nutley, Bohlin, & Klingberg, 2009).

Considering the importance of the proper functioning of attention, because of its involvement in the regulation of thoughts and emotions, maintaining the performance of this process is very important, especially in school, where the child must acquire content in an environment full of countless distractors. The study in question focuses on the attention, proposing and evaluation in the context of the classroom. Thus, it suggests the use of digital games in an integrated way the school activities in the classroom.

The games have features like increasing challenges, rules that establish what can and cannot be done, and involvement of the player in the quest to gain skills and win the game (Kirriemuir & McFarlane, 2004; Prensky, 2006).

We aim to investigate the contributions of the use of a system that integrates cognitive digital games to a database, of the Escola do Cérebro, for monitoring and improvement of cognitive skills, highlighting the attention. The games involve challenges and rules involving the exercise of cognitive functions, especially the working memory, attention and capacity of solving problems.

The study combines qualitative and quantitative approaches. It collects the data based on the observation of the proposed interventions as well as interviews conducted with participating teachers and students to identify their perceptions of digital games’ contributions to the learning process. Furthermore, before and after the implementation of the intervention, we performed a D2 Test of attention that measures selective and sustained attention, as well as visual scanning accuracy and speed.

The intervention consisted in the use of the Escola do Cérebro, using tablets in the classroom, daily for a period of five weeks. The sample consisted of 71 students of the Application School of Basic Education, Federal University of Santa Catarina, aged 7 and 9 years old ($M = 7.64 \pm 1.12$), which were divided into two groups: participant and control. The first ($n = 31$) participated in the intervention, the control group ($n = 40$) was only evaluated using the test before and after the same time interval of interventions.

The Escola do Cérebro is a platform that integrates seven digital games into a database. The application allows visualization of the player's performance and offers the possibility of monitoring by teachers. Students have their scores measured by four variables: time, speed, stability and accuracy.

A statistical analysis was performed based on the application of the paired t-test on the difference of the overall score obtained in the test before and after the intervention in the two groups. The difference in the results obtained from the application of D2 Test of attention before and after was statistically significant (p <0.05), the participant group had mean and standard deviation 60.23 (64.75) respectively, while the control group was 20.00 (42.65).

The result indicates significant improvement in the performance of the sustained attention in the test, as well as a high dispersion, which reveals a variation in relation to the performance.

In addition, students participating in the interview reported a preference for games that involve problem solving, recognize the need to plan actions in relation to their importance for the game and for daily activities, and realize improvements in the ability to sustain attention. The teachers observed changes after the intervention, emphasizing the greater persistence and involvement in school activities, and in some students, improvement in the ability to sustain attention. From this, we conclude that an intervention based on cognitive digital games offers contributions to the learning process and improvement of sustained attention.

Bibliography

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. Acta Psychologica, 129(3), 387–398. http://doi.org/10.1016/j.actpsy.2008.09.005

Dye, M. W. G., & Bavelier, D. (2010). Differential development of visual attention skills in school-age children. Vision Research, 50(4), 452–459. http://doi.org/10.1016/j.surg.2006.10.010.Use

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. Psychol Sci, 18(10), 850–855. http://doi.org/10.1111/j.1467-9280.2007.01990.x

Kirriemuir, J., & McFarlane, A. (2004). Literature Review in Games and Learning. A NESTA Futurelab Research Report, 8, 1–40. Retrieved from https://telearn.archives-ouvertes.fr/hal-00190453/file/kirriemuir-j-2004-r8.pdf

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., & Dehlstrom, K. (2005). Computadorizes training of working memory in children with ADHD-a randomizes, controlled trial. Journal of the American Academy of Child & Adolescent Psychiatry, 44(2), 177–186.

Li, R., Polat, U., Scalzo, F., & Bavelier, D. (2010). Reducing backward masking through action game training. Journal of Vision, 10(14), 1–13. http://doi.org/10.1167/10.14.33

Prensky, M. (2006). “Don’t bother me mom, i’m learning!”: How computer and video games are preparing your kids for twenty-first century success and how you can help! St. Paul, MN: Paragon House.

Rueda, M. R., Checa, P., & Cómbita, L. M. (2012). Enhanced efficiency of the executive attention network after training in preschool children: Immediate changes and effects after two months. Developmental Cognitive Neuroscience, 2(SUPPL. 1), 192–204. http://doi.org/10.1016/j.dcn.2011.09.004

Thorell, L. B., Lindqvist, S., Nutley, S. B., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. Developmental Science, 12(1), 106–113. http://doi.org/10.1111/j.1467-7687.2008.00745.x

Frequently Cited Passages Across Time: New Methods for Studying the Critical Reception of Texts

Jonathan Reeve
jonathan.reeve@columbia.edu
Columbia University, United States of America

Milan Terlunen
m.terlunen@columbia.edu
Columbia University, United States of America

Sierra Eckert
sierra.eckert@columbia.edu
Columbia University, United States of America

Text reuse detection technology and approximate text matching have made possible the large-scale computational identification of intertextuality. These technologies have often been used in plagiarism detection and in studies of journalistic text reuse. Fewer studies, however, have applied these methods to humanities research. We present a method for quantifying the critical reception history of a source text by analyzing the precise location, density and chronology of its citations.

Our work builds on a recent set of digital humanities projects that use text reuse detection in order to study the afterlife of texts through textual quotation. The Viral Texts project and Digital Breadcrumbs of Brothers Grimm take algorithmic approaches to studying text reuse and circulation in their respective fields of research: nineteenth-century popular press and folklore. In both projects, the focus is on using text reuse detection to uncover hidden networks of reuse—the reprinting of short news items and the reuse of motifs or minimal narrative units in folklore—in corpora without standardized conventions of citation. Our project seeks to take this work a step further, by applying a similar method in a different area of cultural production where more standardized citation con- ventions already exist: academic citations. We draw on the work of Dennis Tenen—in using extracted citations to critically visualize the “knowledge domain” of Comparative Literature—and a recent project by JSTOR Labs that uses the text of Shakespeare’s plays and the U.S. Constitution to visualize the scholarship surrounding passages in those sets of texts (Tenen, forthcoming in 2017). Like these projects, we hope to leverage the explicit and institutionalized nature of academic citation in studying text-reuse patterns, using text reuse to ask questions of critical attention in canon formation. By focusing initially on a single text in order to ask about its critical reception history, we hope to provide new methods for studying not only text-reuse patterns, but the sociology of citation practices—studying changes in when, how, and what critics cite from a given text.

In applying these methods to literary scholarship, we've chosen to start on a relatively small scale. George Eliot's novel Middlemarch is an ideal test case, due to its length, copyright status, stable editorial history and canonicity. Perhaps most appropriately for this study, Middlemarch is known for its narrator's wise generalizations—highly quotable fragments that have been featured in more than one Victorian volume of George Eliot's sayings. For the full project we are working with JSTOR Labs to assemble a corpus of all 6,500 articles on Middlemarch in their collection. The figures below display preliminary results with a smaller 483-sample corpus.

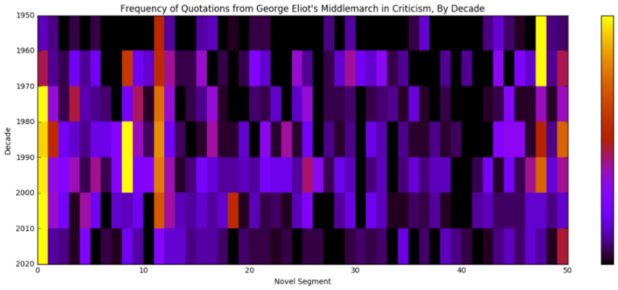
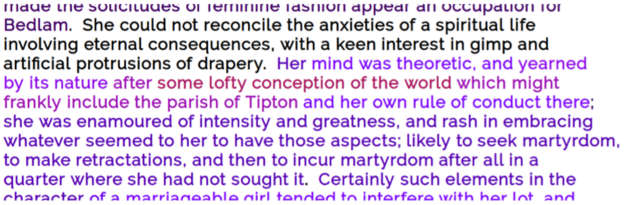


Figure 1. Citation Frequency Heat Map for Middlemarch, by Decade

At the largest scale, Figure 1 shows the frequency of citations across the whole of Middlemarch as a heat map segmented by decade, with yellow signifying the highest citation and black, no citations. Here, the novel is broken into 50 segments along the horizontal axis, and each segment is colored according to the number of times any part of its text has appeared in the critical literature. A number of overall trends are noticeable here. The very beginning and the end of the novel show the most numbers of quotations, followed by the first quarter of the novel. Overall, the second half of the novel, except for the ending, is significantly less quoted from than the first half.

Viewed chronologically, we see that critical interest in certain sections—as expressed in numbers of quotations—appears to shift over time. In the 1950s, most of the critical citation was of the end of the novel, at its climactic third-to-last section, but this interest has faded almost completely by the present day. At the same time, critical interest in the beginning of the novel seemed to be at a relative low point in the 1950s, but quickly became a highly quoted segment by the 1970s. This shift may represent a change in critical attention to the parts of the novel, shifting from the end of the novel at midcentury to the beginning of the novel, where it still appears to rest.



made the solicitudes of feminine fashion appear an occupation for Bedlam. She could not reconcile the anxieties of a spiritual life involving eternal consequences, with a keen interest in gimp and artificial protrusions of drapery. Her mind was theoretic, and yearned by its nature after some lofty conception of the world which might frankly include the parish of Tipton and her own rule of conduct there; she was enamoured of intensity and greatness, and rash in embracing whatever seemed to her to have those aspects; likely to seek martyrdom, to make retractations, and then to incur martyrdom after all in a quarter where she had not sought it. Certainly such elements in the character of a marriageable girl tended to interfere with her lot, and

Figure 2. Citation Frequency Text Browser for Middlemarch

Figure 2 shows a finer grained analysis, no longer segmented by decade, displaying an excerpt from our paragraph-level text browser for Middlemarch. Here again, the color coding represents the number of quotations of each segment, ranging from unquoted black passages, to infrequently quoted purple and red passages and more frequently quoted yellow ones. In our sample, the somewhat abstract contrast between "spiritual life" versus "gimp" and "drapery" remains unquoted, whereas the short, punchy "Her mind was theoretic" has triggered numerous quotations, peaking with the stirring phrase "some lofty conception of the world", before the frequency falls again slightly as the narrator describes how this character's idealism clashes with the real world. This annotated edition could provide both scholars and students of literature with a way to read the novel for passages that have been most discussed in secondary literature, and for passages that have been critically neglected.

The potential applications of this methodology are numerous and wide-ranging. Firstly, this methodology can be used in any discipline to investigate the discipline's theoretical history. As with the 1980 study of Wundt's influence on the field of psychology (Brožek, 1980), our methodology could rapidly and easily produce similar investigations for the influence of Saussure in linguistics, Bourdieu in sociology, Mead in anthropology, Beauvoir in feminist theory, and so on. Moreover, this analysis would be much more fine-grained, registering not only the frequency of works cited but also specifc sections, passages and even key phrases within them.

In addition, we see particular relevance of our methodology to disciplines substantially engaged in questions of quotation and commentary. In theology, commentaries on sacred texts could be analysed in this way to give insights into the texts' interpretive history. In philosophy, the citation of key texts in later periods could be used to assess the shifting priorities of subsequent generations of philosophers. Legal scholars, given the exceptional value placed on precedent in law, have long been interested in using digital technologies and studying citation practices (both of which underpin the recent Ravel project at Stanford University's Law School), which our methodology would make even more granular (not just page numbers but exact phrases).

To return to literary studies, while our own initial project is focusing on academic literary criticism of the post-war period, the methodology could equally be applied to earlier aspects of literary reception. We are interested to examine whether patterns of citation change measurably since English literature is established as a discipline in the late nineteenth century. Going even further back, our methodology could equally be applied to quotations of literary works in non-academic formats, whether non-fiction (newspapers, journals, essays) or other literary works (citations of Shakespeare in Romantic poetry, say).

Our own next steps in this project will be to expand our corpus of quoted texts—the single texts whose citations we are tracing. Do similar patterns of citation apply to Eliot's other novels? Do they apply to novels by Dickens, or by Austen? These analyses will allow us to begin to formulate general hypotheses about patterns of quotation within literary scholarship of recent decades, while also offering unprecedented insights into the works of each of these authors.

Bibliography

Tenen, D. (2017, *forthcoming*) "Digital Displacement," in Futures of Comparative Literature, ed. Ursula K. Heise, Dudley Andrew, Alexander Beecroft, Jessica Berman, David Damrosch, Guillermina De Ferrari, César Domínguez, Barbara Harlow, and Eric Hayot (London: Routledge,).

Brožek, J. (1980). "The Echoes of Wundt's Work in the United States, 1887–1977: A Quantitative Citation Analysis." Psychological Research, vol. 42, no. 1–2, pp. 103–107.

The Tasso in Music Project: A Digital Edition of the Musical Settings of Torquato Tasso's Poetry, ca. 1570–1640

Emiliano Ricciardi
ericciardi@music.umass.edu
UMass Amherst, United States of America

The Tasso in Music Project, currently funded by an NEH Scholarly Editions Grant (2016-2019), is an open-access and interactive digital platform that allows music historians, performers, and literary scholars to access and analyze late 16th- and early 17th-century settings of poetry by Torquato Tasso, arguably the most prominent literary figure of early modern Italy. The project has been realized by a team of scholars from North America and Europe, with the technical support of Stanford University's Center for Computer Assisted Research in the Humanities (CCARH) and UMass Amherst. Upon its completion, the Tasso in Music Project aims to provide upon critical editions of the about 650 extant musical settings of Tasso's poetry, the vast majority of which have never been edited before. These settings represent the work of over 100 composers from a variety of geographic areas and with different musical styles, and as such provide a snapshot of secular vocal music in an age in which it underwent profound transformations. In addition, these settings shed light on Tasso's extraordinary influence on the music of his time– an aspect of his reception that has received surprisingly little attention to date.

The editions, which constitute one of the largest digital repositories of Italian madrigals and related genres, are encoded with Humdrum software tailored specifically for the project by Stanford's CCARH. The editions are presented on the project's website in a variety of electronic formats, such as MEI, MuseData, MusicXML, MIDI, and PDF, among others. They can also be visualized online using Verovio, a recently developed SVG viewer for music encodings. The editions are accompanied by dynamic, in-score critical notes and by music and textual search tools developed by Stanford's CCARH that facilitate analysis of this repertoire. Some of these search tools– such as single pitch and melodic or rhythmic pattern– draw on CCARH's work for the Josquin Research Project, a platform for the analysis of Renaissance music that has received wide recognition in the early music community. Other tools are unique to the project, including those that allow users to run combined musical/textual and vertical sonority searches that are crucial for the study of this particular repertoire.

The platform also features a substantial textual component, with TEI transcriptions of the poetic texts as they appear in the musical settings and in contemporaneous literary sources, both manuscript and printed. The textual apparatus allows for a dynamic visualization of literary variants, thus facilitating the collation of different sources. Thanks to this feature, the project will become an indispensable resource for literary scholars interested in the tradition and transmission of Tasso's poetry, as well as for music historians interested in tracking the literary sources from which composers may have drawn the texts they set to music.

The Tasso in Music Project addresses an interdisciplinary audience, bringing together two institutions invested in the development of digital platforms for musical and humanistic research as well as a group of scholars from the North America and Europe who form the project's editorial and advisory boards. As such, the project may serve as a model for institutional cooperation, opening avenues for interdisciplinary approaches to the creation of digital databases of music and poetry from the late Renaissance and early Baroque periods.

Photogrammar and the Federal Writer's Project: A Model for Teaching Data and Mapping Rhetoric

Courtney Rivard
crivard@email.unc.edu
University of North Carolina Chapel Hill
United States of America

Laura Wexler
laura.wexler@yale.edu
Yale University, United States of America

In a poster session "Photogrammar and the Federal Writer's Project: A Model for Teaching Data and Mapping Rhetoric," we outline the use of some of the educational opportunities we have discovered in using large Digital Humanities projects to create *collaborative partnership pedagogies* to teach students digital rhetoric. In this approach, students become project managers serving the needs of a real world research client – Yale University, the University of North Carolina, Chapel Hill and the University of Richmond's collaborative Photogrammar team.

Bringing together pedagogical theories in Composition and Rhetoric that stress the importance of understanding the role of rhetoric in the construction of digital texts (Sayers, Sample, Galloway) and strategies for equitable collaborative practices in Digital Humanities (Posner and Algee-Hewitt), we teach students to "critically make data" by becoming project leaders tasked with textual mark-up and metadata creation using digitized archival documents. As team leaders, students must think through key rhetorical issues involving naming practices, organizational practices, and user interface and access with a mind toward both historical authenticity and greater inclusion. In this way, they participate robustly in contemporary real world debates about civility, discourse, history and accountability in naming persons, events and things. Therefore, this pedagogical approach opens access into the center of current pedagogical and historical concerns.

The Photogrammar project brings together photographs by the Farm Security Administration taken during the Great Depression and World War II with life histories created through Federal Writers' Project. The new tools currently in development will offer users the opportunity to search and map both the visual and textual histories of the Great Depression in relation to each other. These life histories hold great historical significance as they mark an important precursor to the development of oral history methodology by collecting the histories of people who were previously excluded from the historical record, including women, the working class, and African Americans (Couch, Hirsch, Penkower). Since they will function in some sense as extended captions to the FSA/OWI photographs, adding the life histories also touches upon important issues in the history and criticism of visual culture. Additionally, they hold linguistic keys to racial, gender, class, sexuality and location classification patterns of the time that the photographs alone may not expose. Ethical as well as methodological issues in bringing buried histories to light vividly arise.

Applying collaborative partnership pedagogies in an undergraduate course in Composition, Rhetoric and Digital Literacy at the University of North Carolina that had no prerequisites, students were tasked to create a metadata schema for the life histories to be added to Photogrammar, using TEI to create rules for encoding the text. Because most students entered the class with little technological or collaborative skills, the course places students in small working groups and used Trello (project management software) to organize work flows for lower stake exercises designed to address larger theoretical issues in digital rhetoric (Eyman, Hart-Davidson, Ridolfo). These small stakes activities helped students develop the skills necessary to tackle the larger project that required them to work through critical issues of power stemming from the task of making archival material from the 1930's useful in a digital setting today. For instance, some of the language used to classify racial groups in the FSA and the Federal Workers Project is off-putting, or even offensive today. Students had to come to consensus on such difficult issues because the project required that they present (via skype) their rationale for the metadata schema to a Photogrammar Co-Director. The Co-Director then gave feedback, which students used to write their schema rationale within the TEI of the archival documents. Their decisions did an excellent job of balancing the need to maintain the historical authenticity of the documents while using technology in the service of great inclusion and social justice are now used as the template for the entire life histories collection. Therefore, they not only learned crucial transferable digital skills, but also were given the ability to contribute materially to how the historical record is produced.

Based on the students' work, it will be possible to map the movement of interviewers, trace the prevalence of important issues of the time such as sharecropping, women's labor, WPA work, and mill work, and generate comparative analyzes of rhetoric used via specific interviewer practices. This project highlights the new kinds of classroom as well as scholarly opportunities that arise when rhetorical questions and insights begin to allow students to direct and inflect the further development and build-out process of an established DH site.

Tropy: A Tool for Research Photo Management

Stephen Murray Robertson
srober30@gmu.edu
George Mason University, United States of America

Abigail Mullen
abby@lincolnmullen.com
George Mason University, United States of America

Tropy is a freely licensed and open-source software tool currently under development by the Roy Rosenzweig Center for History and New Media that will allow researchers to collect and organize the digital photographs they take in their research, associate metadata with those images, and export both photographs and metadata to other platforms.

Tropy is filling a critical need in the workflow of researchers who visit archives. This need for photo management has largely arisen since the widespread adoption of photographing sources in the archives instead of reading and analyzing them on-site. Postponing analysis to a later moment in the research cycle proves more difficult than most scholars anticipate: according to the ITHAKA S+R report "Supporting the Changing Research Practices of Historians" (2012), researchers now typically photograph everything they can, and then, faced with "the lack of tools or software to facilitate the process of capturing and using digital photographs for scholars," struggle to "organize and access photographs in a constructive way after a trip." Researchers typically depart from the archives with photographs that might include limited EXIF metadata about the image but never contain any metadata describing the imaged artifact itself. At present it is difficult if not impossible for researchers to attach such metadata to images, particularly regarding their provenance, and to organize them so that scholars can identify and find what is important to their research projects.

With Tropy, you will be able to:

Import. You will be able to drag and drop one or more JPG images into Tropy. Importing will add the image files to Tropy's internal data store, generate thumbnails for each of the new images, and add preliminary metadata based on a template.

Edit. Tropy provides the core functions needed to ensure that images are adequate for your purposes; it is not intended to be a full-featured image editing software. You will be able to rotate, crop, zoom, and adjust contrast. Each image's metadata will also be individually editable. Available fields will be supplied by customizable templates: Tropy will include generalized archive templates, based on Dublin Core and EADS; and researchers and archivists will be able to create their own templates, customized to reflect specific collections and archives. A batch-editing mode will allow users to manipulate metadata across multiple images. Tropy will also include an interface for note-taking and transcription.

Organize. Images will be organized via collections and/or tags, with items able to appear in multiple collections and under multiple tags. You will have many ways of finding your archival images: browsing image collections and tags via list and thumbnail modes; sorting these views using all available metadata, such as date, source archive, and title; and searching across all available metadata, including notes.

Share. All items stored in Tropy will be available for export both locally and to external, web-based services. Exporting a selection of items or a collection from Tropy will generate an archive file that includes image files along with their metadata in machine-readable format. You will also be able to transmit your images and metadata to external services via Tropy plugins. We will create at least three plugins spanning a range of services – Flickr, Omeka and an open-source digital asset management software (DAMS) – as well as documentation that allows users to develop their own plugins.

We anticipate a beta release of Tropy in April 2017, and a 1.0 release in September 2017.

Mnemosyne, Digital Library for Rare and Forgotten Texts (1868–1936): Collections and Digital Editions

Dolores Romero López
dromero@filol.ucm.es
Universidad Complutense de Madrid, Spain

Lucía Cotarelo Esteban
luzia_cotarelo@hotmail.com
Universidad Complutense de Madrid, Spain

José Luis Bueren Gómez-Acebo
Biblioteca Nacional de España, Spain
joseluis.bueren@bne.es

The objective of *Mnemosyne*, Digital Library for Rare and Forgotten Literary Texts (*1868-1936)* is to select, categorize, and make visible (in a digital format) literary texts that belong to a forgotten repertoire in order to allow the historiografical review of the period. *Mnemosyne* has a repertoire of texts and authors who have remained in the shadow of the great literary figures of the first third of the 20th century. This digital library intends to be a field of international experimentation for the creation of interoperable semantic networks through which a large group of scholars could generate innovative research and theoretical reading models for literary texts.

Mnemosyne aims to be an open-access digital library allowing data modelling for **specific collections** (such as intellectual women, Madrid in the Silver Age literature, children kiosk literature, science fiction, photonovels, etc.) in support of research and teaching on Silver Age Spain. Through *Mnemosyne* it can be accessed digital edition of texts. The digitization of these works has been carried out by public and private institutions. The first version of the library is stored on the server of the Universidad Complutense de Madrid Library, itself linked to the collections of the digital library HathiTrust.40 In a search conducted in *HathiTrust* in 2012, the names, surnames, and pseudonyms of selected authors were used to locate a total of 2,873 digitized texts corresponding to "odd and forgotten" writers. The *Biblioteca Digital Hispánica*, which serves as the access portal for the digital collections of the *Biblioteca Nacional de España*, provided 2,448 works by male authors and 1,017 works by female writers. Now we are working on the **interactive editions** of some short stories with the support of the Spanish Biblioteca Nacional.

Behind the scenes of Mnemosyne's public presence online, the project is developing with the aid of the tool *Clavy*. *Clavy* is an RIA (Rich Internet Application) that is able to import, preserve, and edit information from collections of digital objects so as to build bridges between digital repositories and create collections of enriched digital content. *Clavy* also provides a basic system of data visualization, edition, and navigation. There are plans to integrate *@Note*, a collaborative annotation application, into *Clavy*. These two computational tools were developed by the ILSA at the Universidad Complutense de Madrid.44 *Clavy* facilitates the import, export, and edition of records in multiple formats such as MARC2145, as well as their integration into Mnemosine's predesigned model with a view to their export into other compatible formats like XLS (Excel Binary File Format) or XML (Extensible Markup Language), or into other systems like OdA.46 Using *Clavy*, metadata for more than four thousand digitized objects from *HathiTrust* and the Biblioteca Digital Hispánica have already been imported into the *Mnemosyne* database. Logically, the data from these sources was described in MARC21, following the rules for library catalogs. The outcome was forseeable: in some cases, *Mnemosyne's* data model did not require the degree of detail furnished by MARC21, while in other cases it was necessary to incorporate new information absent from that format.

Acknowledgements

Mnemosyne, Digital Library for Rare and Forgotten Literary Texts (1868-1936) is the work of two research teams (LEETHI and LOEP) affiliated with the Universidad Complutense de Madrid and funded by the publicly subsidized national research project "Escritorios Electrónicos para las Literaturas" (Referencia FFI2012-34666, 2012-2015), by the private BBVA Foundation which subsidized the Project "Modelo Unificado de Gestión de Colecciones Digitales con Estructuras Reconfigurables: Aplicación a la Creación de Bibliotecas Digitales Especializadas para Investigación y Docencia" (2015-2017) and by eLITE-CM project "Edición Literaria Electrónica (Ref. S2015/HUM-3426, 2016-2018)".

Information and System Design for Diversity: Can We Do Better?

Amanda Rust
a.rust@northeastern.edu
Northeastern University, United States of America

Julia Flanders
j.flanders@northeastern.edu
Northeastern University, United States of America

Digital information practitioners across many areas seek to preserve and provide access to the voices of disenfranchised and marginalized communities. However, the curation of cultural objects often comes with colonial implications and power-based hierarchical differentials. Providing access to the collections of these groups requires genuine, responsive cooperation, and it also requires that the technical and information systems through which we engage community contributors and participants be equally responsive to diverse cultural circumstances and needs. With support from an IMLS National Forums grant, the Northeastern University Digital Scholarship Group seeks to facilitate a national conversation from which we can learn about responsible partnership in digital projects involving such community-driven collections. We will host a series of public and working meetings to produce a teaching and learning toolkit, to prompt education and change in the approach to systems design for diverse and culturally-sensitive materials.

The need for community-driven pedagogy in this area is well recognized. Practitioners in many fields seek genuine, responsive partnership with the communities in which cultural artifacts were created (See, for example, the repatriation policies of the United States' National Museum of the American Indian or Johnston on recent discussions of repatriation prompted by new publications in the museum field.), deeper understanding of the hegemonic role of knowledge representation via standardized ontological decisions (Berman 1971, Olson 2002.), and investigation into the role of algorithm, interface, and tool design in reinforcing power differentials inherent in the status quo (McPherson 2012, Sweeney 2013, Chun 2011) These topics also arise in venues such as journal special issues (see, for example: *Code4Lib Special Issue on Diversity in Library Technology,* 2015; *Archival Science Special Issue: Keeping Cultures Alive: Archives and Indigenous Human Rights*, 2012; *Library Quarterly Special Issue on Diversity and Library and Information Science Education,* 2013), conference presentations and keynotes (Noble 2015, Matienzo 2015, Cole 2015) such as held by the Society of American Archivists, and current CFPs, such as those of *Digital Humanities Quarterly*, *Archives*, and *ALISE*. New interdisciplinary projects focus on building new methods and tools for archiving media content as a corrective to past silences in the archive (Documenting the Now, Social Media Archives Toolkit, Documenting Ferguson, Our Marathon, to name a few). This work suggests more challenging questions: What ethical decisions inform the dissemination of digital collections? How are naming and representation in our information systems influenced by power? When tools and interfaces guide interactions with documents and items, are those interfaces responsive to community needs, or do they force a diversity of ideas into ill-fitting boxes?

To engage with these issues and inform the toolkit we plan two forum events at which we seek to gather a diversity of perspectives and input on the questions above. The first forum will be held in October 2017, and our preparation for that event will involve a detailed scan of existing LIS and museum pedagogy in the areas central to our focus, and an examination of existing project methodologies and processes. We will also develop a set of design provocations, discussion questions, and a reading list. The poster will present the results of the environmental scan, the design prompts, reading list, and discussion questions for the opening forum, and the preliminary design ideas for the toolkit for critique and commentary by DH2017 conference participants. Our goal for the project is to involve diverse community input at every stage of the design, and this poster represents the first step in that process.

Bibliography

Berman, S. (1971). Prejudices and antipathies: a tract on the LC subject heads concerning people. Metuchen, N.J.: Scarecrow Press.

Chun, W. H. K. (2011). Programmed visions: software and memory. Cambridge, Mass.: MIT Press.

Cole, J. B. (2015, April). Museums, Diversity, and Social Value. Keynote presented at the American Alliance of Museums Annual Meeting, Atlanta, GA. https://aamd.org/our-members/from-the-field/johnnetta-cole-museums-diversity-social-valu

Matienzo, M. (2015, November). To Hell With Good Intentions: Linked Data, Community and the Power to Name. Keynote presented at the LITA Forum, Minneapolis, MN. http://matienzo.org/2016/to-hell-with-good-intentions/

McPherson, T. (2012). "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation." In M. K. Gold (Ed.), Debates in the Digital Humanities (pp. 139–60). Minneapolis, MN: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/29

Noble, S. (October, 2015). Power, privilege and the imperative to act. Invited Keynote at Digital Library Federation conference, Vancouver, Canada. https://open.library.ubc.ca/cIRcle/collections/55474/items/1.0220808

Olson, H. A. (2002). The power to name: locating the limits of subject representation in libraries. Dordrecht, The Netherlands; Boston: Kluwer Academic Publishers.

Sweeney, L. (2013). Discrimination in Online Ad Delivery. Queue, 11(3), 10:10–10:29. http://doi.org/10.1145/2460276.2460278

Stylo: Repenser la chaîne éditoriale numérique pour les revues savantes en sciences humaines

Nicolas Sauret
nicolas.sauret@umontreal.ca
CRC Digital Textualities
Université de Montreál, Canada

Emmanuel Château
emmanuel.chateau@umontreal.ca
CRC Digital Textualities
Université de Montreál, Canada

Arthur Juchereau
arthur.juchereau@umontreal.ca
CRC Digital Textualities
Université de Montreál, Canada

Servanne Monjour
servanne.monjour@wanadoo.fr
CRC Digital Textualities
Université de Montreál, Canada

Marcello Vitali-Rosati
marcello.vitali.rosati@umontreal.ca
CRC Digital Textualities
Université de Montreál, Canada

Michael Sinatra
michael.eberle.sinatra@umontreal.ca
CRC Digital Textualities
Université de Montreál, Canada

Dans cette présentation, nous nous proposons d'analyser la chaîne éditoriale de production et de diffusion des revues savantes en Sciences humaines francophones, en nous appuyant sur le corpus diffusé par la plateforme Érudit. Nous montrerons les limites et les problèmes que présente une telle chaîne éditoriale, pensée à la fin des années 1990, et qui n'a depuis fait l'objet que de très peu de modifications majeures et nous présenterons l'éditeur de texte Stylo, la solution proposée à ces problèmes par l'équipe de la CRC sur les écritures numériques.

Absolument fondamentale en vue d'une diffusion numérique, la structuration des documents et leur balisage sont actuellement reportés à la fin du processus éditorial (l'encodage XML est pris en charge par l'équipe d'Érudit) quand en réalité ceux-ci devraient être envisagés dès la production (par l'auteur lui-même, dès la première révision des articles). À cet égard, il semblerait urgent et logique de réintégrer le travail de balisage et de structuration en début de chaîne.

Un tel objectif se heurte cependant au manque de compétences suffisantes de la part des auteurs : les pratiques d'écriture, telles qu'elles existent aujourd'hui, demandent en effet aux chercheurs de traduire leurs connaissances sur le sens des contenus (ce qu'est une référence ou un titre, l'importance d'un mot, etc.) en simples marques graphiques (italiques, gras, type de guillemets, etc.). Or des erreurs se glissent régulièrement au cours de ce processus du fait du recours à des solutions qui ne sont ni tout à fait spécifiques ni tout à fait explicites, soit finalement à faible valeur de scientificité.

Par ailleurs, ce texte formaté graphiquement est actuellement retravaillé en bout de chaîne par Érudit qui cherche à réinterpréter le sens du texte, à partir de ces indices graphiques. Érudit entreprend alors de transformer ces signes graphiques en balises sémantiques : cette espèce de rétro-programmation implique une perte de temps faramineuse ainsi qu'une potentielle perte d'informations problématique d'un point de vue scientifique.

Pour y remédier, notre équipe est en train de développer l'éditeur de texte Stylo. La philosophie de Stylo consiste à remettre dans les mains des chercheurs la gestion du balisage en partant cette fois-ci de leurs compétences, qui sont des compétences sémantiques plutôt que graphiques. Conçu sur le principe d'un éditeur WYSYWYM (what you see is what you mean), Stylo a le potentiel de changer l'ensemble de la chaîne de production des contenus dans le domaine des revues savantes en Sciences Humaines, en proposant une interface uniformisée et sans perte de données, depuis la rédaction jusqu'à la diffusion, en passant par l'évaluation, la correction et l'édition.

L'éditeur propose aux auteurs un environnement d'écriture très simple mais permettant un enrichissement sémantique à travers des fonctionnalités de balisage léger mais spécifique au travail d'auteur. Le chercheur a ainsi la possibilité de baliser l'ensemble des informations qu'il produit de façon sémantique (titres, sous-titres, références bibliographiques, notes, index), simplement en sélectionnant les éléments, en leur appliquant des styles préprogrammés, ou selon les compétences de l'auteur, en utilisant le format markdown.

La suite de la chaîne éditoriale (évaluation, édition, diffusion) est assurée au sein du même outil, permettant aux producteurs de contenus – les éditeurs de la revue ou de la plateforme de diffusion, ou encore directement les chercheurs eux-mêmes – d'enrichir le texte en sélectionnant et ajoutant des métadonnées (mots-clés, nom de l'auteur, nom de la revue, etc.) et en les alignant avec des autorités.

Les métadonnées, étant alignées, pourront être facilement mises en relation avec d'autres contenus existants (traduction automatique en plusieurs langues, notamment, en se basant par exemple sur les alignements des vocabulaires RAMEAU de la Bnf et LCSH de la Library of Congress). Grâce à l'alignement avec des autorités, les contenus créés via Stylo pourront par ailleurs être mis en relation avec n'importe quels autres contenus – par exemple via un moissonnage qui interroge les autorités, ou encore être mieux indexés par les moteurs de recherche spécialisés.

N'importe quelle plateforme de diffusion peut ainsi exploiter les contenus produits par Stylo. La brique logicielle centrale communiquera via une API avec plusieurs plateformes en fournissant des métadonnées très riches, sans que l'enrichissement soit réalisé par le diffuseur ou l'exploitant final – puisque pris en charge directement par le producteur.

Dans l'esprit des éditeurs WYSIWYM, le formatage graphique du contenu est automatiquement effectué lors de l'export, en s'appuyant sur des modèles programmables et intégrés à la chaîne éditoriale des revues et des plateformes de diffusion (Erudit.ca, Openedition.org).

1. Ces exports pourront être de plusieurs types selon la diffusion visée :
2. des fichiers XML selon les schémas sélectionnés (Erudit, TEI, etc.)
3. des fichiers HTML pour publication directe sur des CMS (grâce à des API)
4. des fichiers print (.pdf) stylés selon des modèles programmables.

L'éditeur pourra être utilisé pour différents types de contenus, notamment :

1. les articles de revues
2. les monographies
3. les mémoires et les thèses
4. les billets de blogue de recherche

Au moment de cette soumission, nous disposons d'un prototype de l'éditeur, dont une version stable est prévue pour le mois d'avril 2017.

Bibliographie

Bachimont, B. (2007). "Nouvelles tendances applicatives : de l'indexation à l'éditorialisation." In L'indexation multimédia. Paris: Hermès.

Doueihi, M. (2011). Digital Cultures. Harvard University Press.

Floridi, L. (2014). The 4th Revolution: How the Infosphere Is Reshaping Human Reality. First edition. New York ; Oxford : Oxford University Press.

Guyot, B. (2004). Sciences de l'information et activité professionnelle. Vol. 38. C.N.R.S. Editions. http://www.cairn.info/resume.php?ID_ARTICLE=HERM_038_0038.

Vitali-Rosati, M. (2015). « An editor for academic papers (xml, html, md, TeX, pdf and if you really need it rtf) », Culture numérique. http://blog.sens-public.org/marcellovitalirosati/an-editor-for-academic-papers-xml-html-md-tex-pdf-and-if-you-really-need-it-rtf/.

HumaReC Project: Digital New Testament and Continuous Data Publishing

Sara Schulthess
sara.schulthess@sib.swiss
Swiss Institute of Bioinformatics, Vital-IT, Switzerland

Anastasia Chasapi
anastasia.chasapi@sib.swiss
Swiss Institute of Bioinformatics, Vital-IT, Switzerland

Ioannis Xenarios
ioannis.xenarios@sib.swiss
Swiss Institute of Bioinformatics, Vital-IT, Switzerland

Martial Sankar
martial.sankar@sib.swiss
Swiss Institute of Bioinformatics, Vital-IT, Switzerland

Claire Clivaz
claire.clivaz@sib.swiss
Swiss Institute of Bioinformatics, Vital-IT, Switzerland

Introduction

HumaReC is a Vital-DH@Vital-IT project funded by the Swiss National Foundation that began on October 1 2016, and will run for two years. The research question of HumaReC is to investigate how Humanities research is reshaped and transformed by the digital rhythm of data production and publication. We use as test-case the edition and the study of a unique trilingual Greek, Latin and Arabic New Testament manuscript known as Marciana Gr. Z. 11 (379). This poster presents intermediate results of our research. For more project details, refer to our open research platform Humarec.org , where HumaReC's research is published.

Change–of–Rhythm in Humanities Research

A two to three years' research project in the humanities has traditionally been characterized by the writing, editing and publication of a printed book. This final step would often be postponed for a certain amount of time after the end of the project. The reasonably delayed publication of printed books has been perceived, in a certain way, as proof of authentic, well done research, certified by an established book series. Recently, however, the digital turn in the humanities is creating a completely new research paradigm by transforming the scheme and rhythm of research through digital writing material. Publishing formats can be videos, draft papers, social media posts, short syntheses of datasets in blogs – all before the research is even completed and peer-reviewed. Certified journals like the New Testament Studies now allow articles to reference blogs of individual scholars (Gathercole, 2015). The peer-review process is becoming a continuous process, rather than a unique event, and is increasingly based on "community-based-filtering" (see Fitzpatrick, 2009) due to the spreading use of open repositories such as HAL for the SSH (Centre pour la Communication Scientifique Directe, 2016) or arXiv for the life sciences (Cornell University, 2016).

"Rhythm" is the key-concept through which we observe the changes happening in DH research. According to Henri Meschonnic's analysis, digital writing reminds one of the presence of orality embedded in any kind of written discourse. Meschonnic considered that writing is not opposite to orality but rather includes it, similarly as sense includes sound (Meschonnic, 1995). For him, orality remains inscribed in writing itself, and this relationship can be expressed by the word "rhythm", since the subject who is speaking always remains related to a performance, to a social act. The speaking subject consists of a "body-social-language", in writing as well as in speaking (Meschonnic, 1982). Thus, rhythm can be used as a key for the mapping of our analysis of the transformations happening in DH: the rhythms of data production, data mining, data editing and publishing; the rhythms of reading, peer-review, echoes and discussions about research.

Continuous publishing as new practice in Humanities

Influenced by the Meschonnic approach, we chose rhythm as a central notion in building the structure of HumaReC: a temporality, based on 24 months, which integrates the diverse levels and parts of the project. A new life sciences journal, *Sciencematters*, is driving a revolution and was a major inspiration. Led by Lawrence Rajendran (Zürich, CH), *Sciencematters* argues that "stories can wait, science cannot" (2016). This outlook allows for the publication of many kinds of data, and fosters the fast publication of small datasets before they are integrated into the "full story" of an article. Following this approach, data relating to our project are published continuously on our open research platform. We want to test two levels of publication in our website, on a small scale:

1) Continuous publishing of small datasets that can be corrected and discussed. This includes short blog articles on current research activities and the publication of new folios of the manuscript (part of HumaReC is a manuscript viewer that displays images and transcription of the three languages). In addition to traditional interactive tools as a forum, we are currently implementing an annotation tool that allows users to comment directly on the manuscript transcriptions.
2) Releases of peer-reviewed fixed data. At four stages that have been preset since the launch of HumaRec, the material will be peer-reviewed by an international board and expanded in a written form best described as a "webbook", as well as an eTalk, a multimedia form of publication that has been developed in Vital-IT.

In addition, we will test different rhythms for transcribing and encoding the manuscript by combining methodologies via automated handwritten text recognition (HTR) using the *Transkribus* tool (Transkribus, 2016).

Our poster will present the data process of the project and the different sections of the platform, listing the research methodologies and the IT challenges of each, as well as their links to Humarec.org.

Bibliography

Centre pour la Communication Scientifique Directe. (2016) HAL. https://hal.archives-ouvertes.fr (Accessed: 20 September 2016)

Cornell University (2016) arXiv. https://arxiv.org (Accessed: 20 September 2016)

Gathercole, S. (2015), "The Gospel of Jesus' Wife: Constructing a Context." *New Testament Studies*, 61(15): 292–313: 10.1017/S0028688515000107

Meschonnic, H. (1982). *Critique du rythme. Anthropologie historique du langage*. Lagrasse: Verdier.

Meschonnic, H. (1995). *Politique du rythme, politique du sujet*. Lagrasse: Verdier.

Sciencematters. (2016) "Stories can wait. Science can't." https://sciencematters.io/why-matters (Accessed: 20 September 2016)

Transkribus Team at the University of Innbruck. (2016). Transkribus. https://transkribus.eu. (Accessed: 20 September 2016)

Vital-DH@Vital-IT (2016) The eTalks: a new digital multimedia editing plaform. https://etalk.vital-it.ch (Accessed: 20 September 2016)

Temporal loci and mixed reality: an experiment in diversifying visualizations of time and space

Celeste Tường Vy Sharpe
csharpe@carleton.edu
Carleton College, United States of America

Sarah Calhoun
scalhoun@carleton.edu
Carleton College, United States of America

Andrew Wilson
awilson@carleton.edu
Carleton College, United States of America

This poster explores ways to use augmented reality to represent complex notions of temporality. Calls for diversifying the digital humanities by scholars like Alan Liu, Amy Earhart, Jessica M. Johnson, and Adeline Koh have called attention to the ways in which digital humanities inquiry and tools often struggle to represent diverse artifacts, cultures, and experiences. Spatial inquiry is one area where scholars are critically engaging and presenting layered analyses of space. Temporality, on the other hand, has received significantly less attention. Notions of time vary widely across cultures. Temporal metadata in digital humanities projects such as timelines or visualizations, however, is frequently constrained by the narrow, linear Gregorian conception of time crystallized by the International Organization for Standardization standard 8601 (ISO 8601).

Using Keith Basso's description of temporal loci of events from his 1996 book *Wisdom Sits in Places*, we can see how myth ("in the beginning"; atemporal) and saga ("modern times"; time described by ISO 8601) frequently overlap and intertwine with each other (50). For instance, in the Thai Buddhist temple paintings that Sandra Cate describes in her 2003 book *Making Merit, Making Art* ("The Defeat of Mara and The Enlightenment (Panya)," plate 10) we can see the convergence of multiple different conceptions of time. One striking piece of a mural captures a moment in time after the recently enlightened Buddha defeats the demon Mara, and a cleansing flood washes away the Mona Lisa and a space shuttle. Attempting to fit the complicated relationships displayed in this image into a simple Dublin Core temporal coverage field would be quite difficult, if not impossible.

We identify two main problems that this initial experiment will address. The first is the issue of visualizing multiple temporalities. Our motivating questions are: what are the visual and spatial relationships between the chronological story of the Buddha defeating Mara given how some Buddhists believe that the Buddha is personal and eternal and always present throughout time? How is that expressed in the mural through a wide range of artistic styles and historical references? These questions will be answered through the course of our research.

The second problem is a more practical question of how to use augmented reality to further research and teaching of these complex cultural concepts when both the visual and technical resources are limited. We intend to use the extant low-res photographs available of the "Defeat of Mara" temple mural and Vuforia to create a cross-platform experience of the religious expression. This will allow users to see and select individual elements in the mural (such as the Mona Lisa or the spaceship) and engage with the different ways one can order and make meaning out of the varied chronologies and temporal references. Vuforia allows us to use an existing framework that has the benefit of being accessible on multiple platforms. We believe this is necessary for facilitating the adoption of augmented reality for classroom and preliminary research uses.

Our poster will outline our theoretical framework, detail our development process using the augmented reality framework Vuforia, and provide possible avenues for further lines of inquiry and applications for temporal visualizations. We'll include static images of the AR experience, as well as ways to access our project remotely.

Long-term outcomes of humanities higher and further education in England and Wales

Nicola Jane Shelton
n.shelton@ucl.ac.uk
University College London, United Kingdom

This paper uses a linked decennial census digital resource (CeLSIUS) from England and Wales supported by funding from the ESRC of interest to the digital humanities community. Publication analysis has shown less use by historians and other humanities researchers, than geographers and other social scientists. This paper introduces the data set and some new research exploring the long-term outcomes of humanities graduates who were resident in England and Wales in 1971 and / or 1991.

Graduates in humanities have lower salaries and lower employment rates in the UK than graduates in medicine and science (ONS, 2013).

But does this affect long term outcomes? The ONS Longitudinal Study for England and Wales (ONS LS) gives us a unique opportunity to study the economic and health outcomes of graduates by university discipline studied. The original sample for the ONS LS was drawn in 1974 (OPCS, 1973) from individuals recorded in the 1971 England and Wales Census. The sample was drawn by selecting four birth dates, giving a sampling fraction of 4/365, or 1.1% of the population of England and Wales, these birth dates are not disclosed. The 1971 sample consisted of around 500,000 people [age 0+], with a similar number of persons (allowing for overall population growth) being sampled at each subsequent census. Sample members are included in all censuses for which they are present and enumerated. More than 200,000 people have been enumerated in five successive censuses (from 1971 to 2011) (Lynch et al, 2015).

In the transition between any two consecutive censuses, some sample members will be lost to the sample either through death or emigration, whilst others will be added to the sample, through birth or immigration. Thus, any child born with an LS sample birth date will automatically become a sample member; similarly someone entering the country (once they enter in to the NHS registration system) with an LS sample birth date will become a sample member. Successful linking clearly depends on the individual being included in the census data capture, and therefore people may effectively leave or enter the record set through enumeration or failure to be enumerated in the census. Blackwell et al (2003) reported tracing rates from 1971 through to 2001 for the ONS LS. These varied from 98.4% in 1991 to 99.3% in 2001; the tracing rate in 2011 was 98.8% (Lynch et al., 2015) which is very high for longitudinal follow up.

All adults with a post age 18 qualification were asked the title of the course, the subject they studied, the year they obtained the qualification and the institution at which they studied, in both the 1971 and 1991 censuses. This is very detailed, offering up to six write in options for subject and has been grouped by ONS into 186 subjects in 1971 and 111 subjects in 1991. Overall participation in higher education in the UK increased from 3.4% in 1950, to 8.4% in 1970, and 19.3% in 1990 (Bolton, 2012). The 1971 respondents will include adults who gained their degrees over their lifetime (so potentially back to the late 1800s). The 1991 graduates will also include anyone with a degree who was enumerated in 1971 and was present and successfully linked in 1991, and also new graduates since 1971 and any immigrants since 1971 who had or obtained a degree prior to the 1991 Census.

In all Censuses occupation (current and for those not employed most recent job and year last worked) is also asked.

In the 1991, 2001 and 2011 Censuses questions about health were asked: general health 2001 onwards and limiting long term illness 1991 onwards. As well as census data, the ONS LS contains linked data on death registrations of sample members.

This paper will examine the employment, health and mortality outcomes for graduates by subject area (humanities, arts, sciences, social sciences) in England and Wales at two time points taking into account age and time period of study. A priori it is expected that those from science disciplines should experience better health outcomes not least due to employment advantages.

Bibliography

Blackwell, L., Lynch, K., Smith, J. and Goldblatt, P. (2003) Longitudinal Study 1971–2001: Completeness of Census Linkage,. Office for National Statistics, London.

Bolton, P. (2012) House of Commons Library Education: Historical statistics Social & General Statistics Standard Note: SN/SG/4252 http://researchbriefings.files.parliament.uk/documents/SN04252/SN04252.pdf

CeLSIUS (n.d.) .University College London. www.ucl.ac.uk/celsius/

Lynch, K., Lieb, S., Warren, J., Rogers, R. and Buxton, J. (2015), Longitudinal Study 2001-2011: Completeness of Census Linkage, Series LS No. 7, Office for National Statistics, Titchfield

ONS (2013) Graduates in the UK Labour Market: 2013 http://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/graduatesintheuklabourmarket/2013-11-19

Collocations and Network Structure as Insights to Functional Elements of Building Adaptive Capacity

Anne R. Siders
siders@stanford.edu
Stanford University, United States of America

Introduction

Decades of field research have identified numerous characteristics and conditions, so-called “determinants,” such as education, wealth, social networks, government transparency, gender, and risk-communication (Brooks et al. 2005; Smit & Wandel 2006; Tol & Yohe 2007), that affect the ability of individuals and groups to prepare for and respond to the effects of climate change. But why these particular determinants and how do they affect our capacity to adapt?

We apply methods from computational text analysis (Sinclair 1991) and network analysis (Brandes 2001; Blondel et al. 2008) to offer an innovative approach to understanding the concept of adaptive capacity. Computational analyses allow us to reveal the unconscious rhetoric of adaptive capacity that illuminates how determinants are interconnected and how they might work to build adaptive capacity. These patterns are not stated or visible in a single study but emerge from a field-wide analysis. Results depict a concept map of adaptive capacity that operationalize the academic discourse and highlights points of convergence and divergence to inform future research.

Corpus and Methods

A Web of Science search for title = “adaptive capacity”, years 1800-2015, returned 448 non-duplicate English language academic articles. Based on title, journal, and abstract, we categorized papers as focused on social (e.g., community, organization, government; n=295) or non-social systems (e.g., biological, engineering; n= 153). Of 295 social papers, 261 full length texts (88%) were accessible. Most (91%) were published post-2001, when the Intergovernmental Panel on Climate Change (IPCC) first recognized adaptive capacity as a major element of vulnerability to climate change (IPCC 2001), signaling the concept’s rise to the forefront of climate and sustainability research.

We used collocation analysis to develop a network of determinants that visualizes inter-connections and may be interrogated. Based on a close reading, we identified 164 determinants of adaptive capacity and 351 related terms (to account for regional spelling variations, synonyms, gerunds, etc.). Collocation analysis was used under the theory that two concepts whose terms frequently co-locate have a conceptual relationship. Collocates were identified in symmetric 15 word distance with significance of 0.01 using a Fisher’s Exact Test.

Measures of network structure, such as centrality and modularity, have been found in other fields to provide insights into functionality (Krackhardt 1990; Danon et al. 2005). Collocations between determinants were visualized as a network (149 nodes, 1877 edges, network density 0.09). Both degree and betweeness centrality (Brandes 2001) were calculated. The centrality of a determinant may provide insight as to its role and sphere of influence. Community detection (Blondel et al., 2008), which has been shown in other cases to reveal functional groups (Danon et al. 2005), was performed 10 times each at three resolutions (0.4, 0.7, 1.1) (Lambiotte et al. 2008).

Results

Results provide substantial insight into potential roles for determinants. In many cases, results confirm expectations and establish consensus. In others, results raise new research questions and may provide an impetus to test assumptions currently held within the field. Results further suggest determinants group into hierarchical functional modules, which could provide a function-based framework to assess adaptive capacity. These patterns may also reconcile competing theories in adaptive capacity literature as to whether all determinants are critical or some may compensate for weaknesses in others (Tol & Yohe 2007).

Bibliography

Blondel, V.D. et al., (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics,* (October), pp.1–12.

Brandes, U., (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2), pp.163–177.

Brooks, N., Adger, W.N. & Kelly, P.M., (2005). The determinants of vulnerability and adaptive capacity at the national level and the implications for adaptation. *Global Environmental Change*, 15(2), pp.151–163.

Danon, L. et al., (2005). Comparing community structure identification. *Journal of Statistical Mechanics,* 2005(9), p.P09008.

IPCC, (2001). *Climate Change 2001: Third Assessment Report, Contribution of Working Group II: Impacts, Adaptation and Vulnerability* J. J. McCarthy et al., eds., Cambridge University Press.

Krackhardt, D., (1990). Assessing the political landscape: Structure, cognition and power in organizations. *Administrative Science Quarterly,* 35(2), pp.342–369.

Lambiotte, R., Delvenne, J.-C. & Barahona, M., (2008) *Laplacian Dynamics and Multiscale Modular Structure in Networks.* arXiv:0812.1770.

Sinclair, J., (1991). *Corpus, Concordance, Collocation,* Oxford, UK: Oxford University Press.

Smit, B. & Wandel, J., (2006). Adaptation, adaptive capacity and vulnerability. *Global Environmental Change,* 16(3), pp.282–292.

Tol, R.S.J. & Yohe, G.W., (2007). The weakest link hypothesis for adaptive capacity: An empirical test. *Global Environmental Change,* 17(2), pp.218–227.

L'accès vu comme une contrainte et un défi : le processus de naissance d'une base de données

Pinelopi Skarsouli
pinelopi.skarsouli@cnrs.fr
National Center for Scientific Research (CNRS), France

Dina Bacalexi
dina.bacalexi@cnrs.fr
National Center for Scientific Research (CNRS), France

Une base de données organise la connaissance en vue de son utilisation par un public donné. Cette définition trop rapide et lacunaire ne tient pas compte de la complexité du processus de création, ni de la véritable nature d'une base numérique, qui, pour nous, doit être le fruit d'un dialogue entre la partie scientifique et la partie technique du projet. Nous allons étudier, pour deux bases antiquisantes, le processus de passage d'une approche textuelle, libre de toute contrainte, à une autre basée sur la structuration, c'est-à-dire la transcription en structuration formelle des décisions prises par les analystes des textes antiques.

Dans ces deux cas, un nouvel accès, résultat de la restructuration des outils, vise un public nouveau tout en servant mieux l'ancien :

1) l'abandon d'un modèle conçu pour l'imprimé, où la recension annuelle des données pour produire le volume papier de *l'Année Philologique*, accompagné d'une transformation en site web, constituait pendant des décennies le but recherché : l'objectif était un changement total du contenu et de la structure, pour obtenir une véritable base de données, tout à fait différente (*IPhiS,* la nouvelle base philologique en construction).
2) la transformation d'un modèle sans volume papier paru, qui n'était pas pour autant une construction « native numérique » mais un simple export html périodique d'un ensemble de données thématiques : l'objectif était la reprise des anciennes données avec une modélisation nouvelle permettant d'obtenir une nouvelle base, qui continuerait l'ancienne seulement en ce qui concerne le contenu (le cas du *Répertoire des sources philosophiques antiques*).

La question de l'accès libre aux données s'est posée à nous comme une contrainte. Notre travail d'antiquisants analystes de sources grecques et latines philologiques ou philosophiques (au sens large des termes) a croisé celui des développeurs d'applications numériques quand nous avons commencé à transformer les deux outils bibliographiques précités. Il existe plusieurs façons d'envisager la perméabilité et l'articulation scientifique-technique ; nous étudierons comment le dialogue permanent direct permet un meilleur accès aux données d'une base.

Quand un modèle a fait ses preuves depuis près de 80 ans et que sa notoriété est installée dans la communauté, la difficulté est d'imposer un modèle nouveau quant à la pertinence scientifique et le traitement de l'information. La contrainte est double : passer d'un logiciel propriétaire (4D) dont l'adaptation convenait tout à fait à la bibliographie papier et à son export en ligne, à une solution de logiciel libre qui apporterait cependant la même richesse et finesse d'indexation que 4D et créerait une véritable base de données numérique. La question ne concernait pas seulement le développement de l'outil nouveau, mais aussi le type des données à implémenter dans la base. Notre questionnement concernait aussi le fait même de rendre la base accessible librement sur le web : au moment de la prolifération de l'information grâce aux moteurs de recherche généralistes du type *Google scholar*, quelle serait la « plus-value » d'une énième base ? C'est pourquoi nous avons commencé à réfléchir comment les entrées de la base pourraient pointer vers des données textuelles littéraires extérieures en XML/TEI, afin que nos bases ne soient pas simplement des *thesauri* bibliographiques riches et bien ordonnées, mais sans lien avec les sources qu'elles sont censées analyser. Cette question est pour l'instant pour nous au stade de la réflexion théorique.

La liberté d'accès, dans le cadre des disciplines rares comme les nôtres, pose la question de la protection contre le copiage, le plagiat, l'imitation et donc l'appropriation de notre travail. C'est pourquoi, dès le début du projet, il est important de choisir une licence adaptée, par exemple de la famille *Creative Commons.* Il faut aussi envisager l'*accessibilité effective* des données *via* l'interface utilisateur et le respect des standards du web.

Nous constatons trop souvent que les bases de données, sites collaboratifs, plateformes, éditions etc. émanent de programmes limités dans le temps, au financement et au personnel précaire, ce qui impacte la réalisation des projets dans les délais impartis. Or, si ces bases répondent à une demande d'accès à des données fiables et de qualité contrairement à la masse incontrôlée offerte par les moteurs généralistes du web, elles ont vocation à durer et ont besoin d'une mise à jour régulière de leur contenu et de leur infrastructure logicielle pour remplir leur mission. Les changements de technologie (internes ou imposés par l'écosystème) peuvent comporter des risques pour les projets et leurs utilisateurs finaux. La disparition des outils entraîne souvent celle du savoir et du savoir-faire des personnels.

La nécessité d'obéir aux injonctions du court terme impacte le travail de renouvellement des anciennes bases : les délais de transformation sont plus longs que prévu, puisqu'il faut d'abord passer par une expression de nouveaux besoins et ensuite par une modélisation tenant compte des choix scientifiques et traduisant ceux-ci en solutions techniques ; il faut ensuite expérimenter, adapter, affiner. Un tel fonctionnement est à l'opposé d'une demande de solution rapide clés en main.

L'implémentation des notices dans une base en flux continu sur le web (en vue de leur moissonnage par les moteurs de recherche en temps réel), le rêve d'exhaustivité, contribuent à alimenter l'idée de liberté totale individuelle qui va à l'encontre du mode de travail collectif. Le risque est d'instaurer une nouvelle forme de dépendance et de contrôle généralisé, voire d'accélération des cadences et de fin programmée de toute vérification de la pertinence scientifique. C'est tout à fait différent d'une base de données participative et basée sur la contribution de la communauté.

Le travail sur le contenu d'une base dépend aussi de l'élaboration d'un cahier de charges, de la modélisation des données (qui définit à son tour le type d'accès aux données dans l'immédiat mais aussi dans l'avenir), des formulaires permettant de la remplir, de son interface de consultation et de la constitution d'une communauté d'utilisateurs. La question est la (re)structuration et la (re)présentation des données sous une nouvelle forme en ligne, qui tienne compte des standards actuels (libre accès, interopérabilité etc.). La normalisation des données et la réduction de la redondance de l'information sont des gages d'une base de données pleinement accessible et exploitable par ses utilisateurs.

Bibliography

Bénel, A. 2014. "Quelle interdisciplinarité pour les 'Humanités numériques'?" *Les Cahiers du numérique* 10/4: 103-132. http://www.cairn.info/revue-les-cahiers-du-numerique-2014-4-page-103.htm

Berra, A. 2015. "Pour une histoire des humanités numériques", *Critique* 8 (n° 819-820): 613-626. http://www.cairn.info/revue-critique-2015-8-page-613.htm

Burnard, L. 2012. "Du *literary and linguistic computing* aux *digital humanities*: retour sur 40 ans de relations entre sciences humaines et informatique" in *Read/Write Book: une introduction aux humanités numériques*, Pierre Mounier (dir.), Marseille. http://books.openedition.org/oep/242

Clivaz, Cl., Dilley, P. and Hamidovic, D. ed. 2016. (in collaboration with A. Thromas), *Ancient Worlds in Digital Culture*, Leiden: Boston.

Crane, G., Seales, B. and Terras, M. 2009. "Cyberinfrastructure for Classical Philology". *Digital Humanities Quarterly* 3/1 www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html.

Flanders, J. 2009. "The Productive Unease of 21st-century Digital Scholarship." *Digital Humanities Quarterly* 3/3 www.digitalhumanities.org/dhq/vol/3/3/000055/000055.html.

Le Deuff, O. ed. 2014. *Le temps des humanités digitales. La mutation des sciences humaines et sociales*, Limoges.

Modeling Student Authorship: The Rhetoric of Markup in the Writing Classroom

Kevin G. Smith
kgsmith2@gmail.com
Northeastern University, United States of America

This poster reports on research that examines the use of XML markup for student authoring, a marked shift from the mimetic roots of XML and its primary use in digital humanities research, the TEI. Thus far, two courses have been taught with a total enrollment of 35 students (19 students in advanced writing for the technical professions, Summer 2016; 14 students in first-year writing, Fall 2016). The research questions are: How does markup function rhetorically when used for authorship? Does writing in XML and designing schemas for authoring contribute to students' understanding of their writing and reading processes? Do reading and writing practices in the markup classroom transfer to other contexts?

These research questions present unique methodological concerns for the study of markup. How can we make claims about the rhetorical and expressive capacities of *authorial* markup? How can we better understand the role of the schema, the markup, and the platform(s) in students' writing, reading, and thinking processes? In short, how do we study this? These questions will animate the presentation of preliminary results, the subject of the poster.

There has been considerable interest in the semantics of markup languages at recent TEI (Ciotti & Tomasi, 2014; Eide, 2013) and DH conferences (Sperberg-McQueen Marcoux, & Huitfeldt, 2010, 2014), most of which has centered around formal approaches to modeling semantics to explicate the meaning of markup (Sperberg-McQueen, Huitfeldt, & Renear, 2000). A related thread of markup research has examined the rhetorical and expressive capacities of markup (Flanders, 2004; Flanders & Fiormonte, 2007), growing out of an understanding of markup as not merely descriptive, but also interpretive and, indeed, performative (Renear, 2000). Though Wernimont and Flanders (2007) have discussed the potentials of authorial markup to expand our shared notions of scholarly communication, markup in this authorial realm remains rarely used (one exception being the work of Desmet et al., 2005) and even more rarely studied in a systematic way. This poster will present preliminary results from just such an attempt—a sustained study of an experimental approach to XML as a technology for the production of texts.

The production of texts, in this case, was undertaken by two cohorts of undergraduate students. In addition to writing their assignments in XML (using Oxygen), these courses engaged students in a semester-long, collaborative writing project: the design and implementation of a custom XML schema that structurally and rhetorically models a range of genres of writing. Pedagogically, this approach aims to foster the close attention and metacognition often cited in classroom-oriented uses of XML/TEI (e.g., Singer, 2013; Conatser, 2013). Where this approach to markup differs from earlier uses, however, is in the thoroughly bottom-up, data driven approach to schema design (Piez, 2001). Students begin with a (basically) bare schema and—iteratively and deliberately over the course of an entire semester—design and revise the schema for a range of writing tasks using document analysis and modeling, qualitative writing research methods, and their own experiences of authorship.

To research these classes, I employed a teacher research methodology—a systematic approach to data collection that **honors the inside perspectives of teachers and students**—that adapted qualitative research methods culled from ethnography, education, and writing studies research. Data was gathered from direct participant observation, reflective journaling, qualitative interviews (three interviews each with nine case study students), survey, and the collection of student writing (normal prose and XML, including version control logs for all XML files). Teacher research foregrounds and honors the experiences and perspectives of students as they compose; thus, the particular methods deployed in this study concern writing as a process, rather than as a static product. This methodology aligns with research in rhetoric and writing studies, which, fundamentally, understands "writing (and broader rhetorical practice) as a verb rather than a noun" (McNely and Teston, 2015: 115).

Preliminary results from the study speak to 1) how students develop and operationalize genre knowledge; 2) the rhetorical constraints and affordances of schema design as collaborative writing; and 3) markup's reported intervention into students' thinking and writing processes.

This poster frames this research study as a case study, a demonstration of the kinds of insights that systematic, qualitative research into markup can foster. It aims to organize the audience around a series of questions, including: How are rhetorical theories pertinent to our examination of DH tools and methods, particularly those of data modeling and representation? How does the study of student writing/authorship necessitate a willingness to invent methods sensitive to, and emergent from, particular sites of research? What methodological (re)orientation does an expansion of our disciplinary objects of inquiry require? These questions are best explored interactively, through the dynamic presentation of data generated through this research, and an exploration of the opportunities and limitations of this qualitative approach to markup research.

Bibliography

Ciotti, F., Lana, M., & Tomasi, F. (2014). "TEI, ontologies, linked open data: Geolat and beyond." *Journal of the Text Encoding Initiative*, (8).

Conatser, T. (2013). "Changing medium, transforming composition." *Journal of Digital Humanities*, *2*.

Desmet, C., Balthazor, R., Cummings, R., Hilton, N., Mitchell, A., & Hart, A. (2005). "<emma>: Re-forming composition with XML." *Literary and Linguistic Computing*, *20*(Suppl): 25-46.

Eide, Ø. (2014). "Ontologies, data modeling, and TEI." *Journal of the Text Encoding Initiative*, (8).

Flanders, J. (2004). "The rhetoric of performative markup." *Critical Inquiry*, *31*(1): 49-84.

Flanders, J., & Fiormonte, D. (2007). "Markup and the digital paratext." *Digital Humanities 2007: Conference Abstracts*.

McNely , B., & Teston, C. (2015). "Tactical and strategic: Qualitative approaches to the digital humanities." In Jim Ridolfo and William Hart-Davidson (eds.), *Rhetoric and the Digital Humanities*. Chicago: University of Chicago Press, pp. 111-126.

Piez, W. (2001). "Beyond the 'descriptive vs. procedural' distinction." *Markup Languages*, *3*(2): 141-172.

Renear, A. (2000). "The descriptive/procedural distinction is flawed." *Markup Languages*, *2*(4): 411-420.

Sperberg-McQueen, C. M., Huitfeldt, C., & Renear, A. (2000). "Meaning and interpretation of markup." *Markup Languages,* 2(3): 215–34.

Sperberg-McQueen, C. M., Marcoux, Y., & Huitfeldt, C. (2014). "Transcriptional implicature: A contribution to markup semantics." *Digital Humanities 2014: Conference Abstracts*.

Sperberg-McQueen, C. M., Marcoux, Y., & Huitfeldt, C. (2010). "Two representations of the semantics of TEI Lite." *Digital Humanitis 2010: Conference Abstracts*.

Wernimont, J., & Flanders, J. (2011). "Possible worlds: Authorial markup and digital scholarship." *Digital Humanties 2011: Conference Abstracts*.

The VSim Repository and Archive: Knowledge Mobilization for 3D Research

Lisa M. Snyder
lms@ats.ucla.edu
UC Los Angeles, United States of America

With the explosion of academic interest in 3D modeling and interactive environments, two questions dominate: How can scholars and educators access these models? And how can they be integrated into the classroom? Enter VSim – a much-needed software interface for academics exploring applications for 3D computer models in humanities research and teaching – and the VSim Repository and Archive, both now available through the UCLA Library.

VSim is an easy-to-use software interface that allows real-time exploration of highly detailed, academically generated 3D computer models in both formal and informal educational settings across grade levels and humanities disciplines. The VSim Repository and Archive provides a platform for the dissemination of academically generated 3D content.

This poster presentation is both a launch announcement and a solicitation for submissions from content creators. If you are working in 3D, please consider sharing your projects for teaching and learning through the VSim Repository and Archive.

Ghostwriter identification in Yasunari Kawabata's works in the 1960s

Hao Sun
sonnkou1985@gmail.com
Doshisha University, Japan

Mingzhe Jin
mjin@mail.doshisha.ac.jp
Doshisha University, Japan

Introduction

Yasunari Kawabata was a Japanese novelist who received the Nobel Prize for Literature in 1968. He was famous for his masterpieces such as *Snow Country*, *The Sound of the Mountain*, *The Old Capital*, *House of the Sleeping Beauties,* and so on. Kawabata had a shattered childhood. He was orphaned at five years old, and his other relatives including his grandfather, grandmother, and elder sister also passed away before he was fourteen. The successive deaths of his loved ones induced mental problems, which became worse in the 1960s. He became addicted to sleeping pills during those years. However, two of Kawabata's masterpieces *The Old Capital* and *House of the Sleeping Beauties* were published during his sleeping pills addiction period. These two novels were suspected as having been written by ghostwriters because it was hard to imagine that Kawabata could continue writing novels in his mental condition. There are already some pieces of evidence for the ghostwriter issue of *The Old Capital* and *House of the Sleeping Beauties*. Kawabata sent a letter to Sawano before the publication of *The Old Capital*. In the letter he wrote: "I have accepted to write a novel about Kyoto; the deadline is approaching but I even don't know how to start. Sawano was surprised when he received this letter. He went to Kyoto to meet Kawabata and gave advice on how to write *The Old Capital*. Itasaka mentions in his book that *The Old Capital* and *House of the Sleeping Beauties* were actually written by ghostwriters (Itasaka, 1997). *The Old Capital* may have been written by Kawabata's three disciples whose names are Hisao Sawano, Makoto Hokujyo, and Yukio Mishima. *House of the Sleeping Beauties* may have been written by Yukio Mishima. In this study, we show strong evidence suggesting the real author of *The Old Capital* and *House of the Sleeping Beauties* from a data analysis approach.

Method

The method of this study includes three main steps. Firstly, we digitized more than ten novels of both Kawabata and the three possible ghostwriters. Then, we extracted stylometric features from the novels, and all chapters of *The Old Capital* and *House of the Sleeping Beauties*. Finally, we applied the unsupervised and supervised methods to infer the possible author of *The Old Capital* and *House of the Sleeping Beauties*.

We used bigrams of characters and punctuation marks, part-of-speech bigrams, and phrase patterns as stylometric features, which have been proven useful in Japanese authorship attribution (Matsuura and Kanada, 2000; Jin, 2003, 2013).

Bigrams of characters and punctuation marks are pairs of two adjacent characters or punctuation marks extracted from plain text. Japanese texts should be tokenized previously for the extraction of part-of-speech features. We applied the Japanese morphological analyzer called MeCab to separate a Japanese sentence into morphemes. MeCab outputs the parts-of-speech of words in several layers. Deeper layers process more detailed information. In this study, we use information from the first layer. As an example, part-of-speech bigrams in the Japanese sentence “Ronbun wo kaku.” are “Noun_Particle,” “Particle_Verb,” and “Verb_Period.”

Phrase pattern is a powerful feature that can be extracted in terms of syntax. A Japanese parser (CaboCha) was introduced to separate sentences into phrases. Phrase pattern is defined as the smallest unit that divides the sentence into unnatural parts (Jin, 2013). It is a combination of two parts. One is the original form of the particles and punctuation marks, while the other is the parts-of-speech of the other materials, except for the particles and punctuation marks in the same phrase. The two phrase patterns in the sentence “Ronbun wo kaku.” are “Noun_wo” and “Verb_Period.”

We applied unsupervised methods and the integrated classification algorithm in this study. The idea of the integrated classification algorithm was to combine the results of several stylometric features and classifiers. It achieved highest classification accuracy in authorship attribution of literature (Jin, 2014). The integrated classification algorithm combines the results of stylometric features and classifiers to avoid the bias under a majority vote rule. AdaBoost (ADA), High-dimensional Discriminant Analysis (HDDA), Logistic Model Tree (LMT), Random Forest (RF), and Support Vector Machine (SVM) were used as the base classifiers.

Results

The result of *House of the Sleeping Beauties* reveals that compared to Mishima, all chapters of *House of the Sleeping Beauties* are more likely to be written by Kawabata. The result in the classification between Kawabata and Hokujyo shows that the writing style in all chapters of *The Old Capital* is more like Kawabata’s.

Bibliography

Itasaka G. (1997*). Gokusetsu Mishimayukio-Seppuku to furamenko.* Natsumesyobo Press.

Matsuura, T. and Kanada, Y. (2000). Identifying Authors of Sentences in Japanese Modern Novels via Distribution of N-grams. *Mathematical linguistics,* 22: 225-238.

Jin, M. (2003). Authorship Attribution and Feature Analysis Using Frequency of JOSHI with SOM. *Mathematical linguistics*, 23: 369-386.

Jin, M. (2013). Authorship Identification Based on Phrase Patterns*. The Japanese Journal of Behaviormetrics*, **40:** 17-28.

Jin, M. (2014). Using Integrated classification Algorithm to Identify a Text's Author. *The Japanese Journal of Behaviormetrics*, 41**:** 35-46.

Access to DH Pedagogy as the Norm: Introducing Students to DH Methods Across the Curriculum and at a Distance

Dan Tracy
dtracy@illinois.edu
University of Illinois at Urbana-Champaign
United States of America

Elizabeth Massa Hoiem
hoiem@illinois.edu
University of Illinois at Urbana-Champaign
United States of America

This poster presents research into integration and assessment of digital humanities pedagogy in a distance course on the History of Children’s Literature, and provokes conversation about pedagogical approaches that expand student access to DH methods, tools, and dispositions. Much of the existing literature on DH pedagogy addresses methods courses or multimodal writing courses rather than integration of DH practices in particular topical contexts, or advanced topics courses that explore a narrow slice of disciplinary content through extended engagements with digital projects (Ball 2012; Mostern & Gainor 2013; Fyfe 2016; Nyhan, Mahony, and Terras 2016). This literature provides valuable lessons but raises questions about the feasibility of engaging with DH across the curriculum in small-to-medium scale engagements with new methods and technologies. Amy E. Earhart and Toniesha L. Taylor (2016), for example, respond to this situation by rejecting the idea that DH should be limited to advanced courses and propose broader integration of “embedded [DH] skills development” that students can take out of the environment of a specific institution. Similarly, we suggest that allowing for repeated and diverse engagement by students across methods-intensive and topic-intensive courses (as is now common for writing) is necessary for teaching deeper DH dispositions like collaboration, openness to failure, and creativity with technology.

Simultaneously, the existing literature has focused on residential instruction with access to physical artifacts. This limit is problematic when at least one discipline with a heavy investment in DH, library and information science, is well past transition to a majority distance learning population. LIS programs have developed experience and expertise in teaching technology at a distance, and lessons from these programs may be useful to the DH community. While some teaching goals may only be met in person, others might be achieved through well-structured online learning.

To ground this discussion, the authors, the course instructor, and a subject librarian will present their development, assessment, and rethinking of a multimodal publication assignment using the Scalar platform in a synchronous online course on the History of Children's Literature. Students worked in groups to create a multi-media web resource on "diverse history." The class discussed what is included or omitted from historical narratives, whether they be children's historical fiction or history textbooks, before contemplating this selection process in children's literature itself. The librarian introduced students to the context of DH publishing and Scalar, and to issues related to responsible use of multimedia. Then each group chose an issue related to "diverse history" and built one section of the website. The long-term goal is for successive classes to edit, revise, and expand this project

This collaborative project replaced an assignment from previous years, when students built individual websites about a children's book of their choice. This project maximized scaffolding, with detailed guidance on information students should locate about their books and the final website shape. This iteration of the class took place during a time when distance students came to campus one weekend each semester, and this time was used for in-depth introduction to the array of specialized library resources needed to complete the questions about their book's production and reception. The new assignment sought to re-imagine learning outcomes that would allow students to engage with a particular DH publishing technology, Scalar, and grapple with issues of collaboration and multimodal authoring in a context where the final product was less predetermined. Nonetheless, the elimination of the in person component, which occurred at the same time, removed an obvious "lab" opportunity for learning related technical issues. The pedagogical design involved making the best balance between asynchronous and synchronous activities to compensate for the absence of in person activities. Our evaluation of the success of the assignment relied on assessment of Scalar sample sites and final projects created by the students, as well as on reflective essays written by the students and observations made in the course of student consultations. This evaluation led to ideas for how to revise the course for future semesters to improve learning of collaborative behaviors, openness to failure, and creativity with technology. This includes, most notably, a re-envisioning of how synchronous class time is used in the future.

By sharing our experiences in developing, teaching, assessing, and revising this course in successive iterations, we hope to explore with attendees the ways in which DH methods, tools, and dispositions can proliferate across the curriculum. We will promote discussion of what DH methods, tools, and dispositions can be taught well in different settings, whether that means varying scales of integration in DH classrooms, or exploring what can be taught virtually versus in person.

Bibliography

Ball, C. E. (2012). "Assessing Scholarly Multimedia: A Rhetorical Genre Studies Approach." *Technical Communication Quarterly* 21: 61–77.

Earhart, A. E., and Taylor, T L. (2016) "Pedagogies of Race: Digital Humanities in the Age of Ferguson." *Debates in the Digital Humanities 2016*. Ed. Matthew K. Gold and Lauren F. Klein. U of Minnesota P, Minneapolis. 251–64.

Fyfe, P. (2016). "Mid-Sized Digital Pedagogy." *Debates in the Digital Humanities 2016*. Ed. Lauren F. Klein and Matthew K. Gold. University of Minnesota Press. 104-117.

Mostern, R., and Gainor, E.. (2013). "Traveling the Silk Road on a Virtual Globe: Pedagogy, Technology and Evaluation for Spatial History." *Digital Humanities Quarterly* 7.

Nyhan, J., Mahony, S., and Terras, M. (2015)"Digital Humanities and Integrative Learning." *Integrative Learning*. Ed. Daniel Blackshields, James Cronin, Bettie Higgs, Shane Kilcommins, Marian McCarthy, and Anthony Ryan. London: Routledge. 235-47.

How Agatha Christie Described Women? : The Behaviour of *She* in Christie's Novels

Narumi Tsuchimura
t.naru.425@gmail.com
Osaka University, Japan

This paper describes a stylometric analysis of Agatha Christie's works with a special reference to the use of the feminine personal pronoun *she* in her novels. In a previous study, Tsuchimura (2016), as a result of the statistical analysis, it is shown that Christie tends to use the word *she* much more frequently than her contemporary Dorothy Sayers, a British mystery writer. She occurs about 12,000 times per million words in Christie's works whereas it occurs about 5,500 times in Sayers' works. A number of characteristic words were identified in the study, and this study focuses on the use of the word *she* in collocations within Christie's works.

It is possible to hypothesize that *she* occurs frequently in Christie's works because although the protagonists in Sayers' works are all male (Lord Peter Wimsey), Christie frequently employs female protagonists (Miss Marple and Tuppence Beresford). In order to test this hypothesis, a Random Forest (Breiman, 2001) classifier is trained on the 500 most common words from all of Sayers' works (55 texts) and Christie's works whose protagonists are not females (173 texts). As is shown in Figure 1, looking at the mean decrease in the GINI importance of the model per word, which measures the relative importance of each word in classifying a text as that of Sayers' or Christie's, the word *she* contributes strongly to classification of texts into

2 groups regardless of the protagonists' gender. We can see that Christie tends to use the word *she* frequently even in the works having or foregrounding male protagonists.

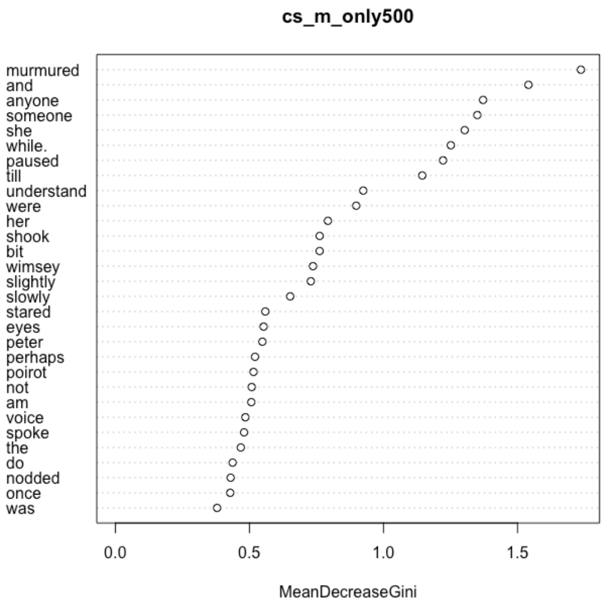


Figure 1. Variable Importance Plot of the result of Random Forests

The question arises as to how the word *she* is used in Christie's novels. To clarify its behaviour, this study examines collocations containing *she* in Christie's works. Following Sinclair and Jones (1974), who state that 'for any node, a very high proportion of relevant information could be obtained by examining collocates at positions N-4 to N+4,' this study deals only with collocates at position N-4 to N+4. From a number of statistical measures commonly used to calculate the significance of collocations, this study chooses the MI-score, for 'MI-score tends to give information about its lexical behaviour, but particularly about the more idiomatic ('fixed') co-occurrences' (Hunston, 2002: 74). The minimum frequency of collocates was set at 10, and the threshold of the MI-score was set at 3.0.

The extracted collocates of *she* in Christie's works amount to 104, when proper nouns are excluded, the number of remaining types of collocates is 79. When compared with collocates of *she* in Sayers' works, those in Sayers' novels are mostly function words. It is thus difficult to see the behaviour of *she* in Sayers' works and in comparison with Christie's works.

When we compare collocates of *she* with those of *he* in Christie's works, we can see stereotypical femininity. The collocates of *she* in Christie works consist of a large number of words related to crying (*sobbed, wept, wailed, screamed, crying*) and actions indicative of fear or sadness (*prostrated, frightens, fainted, shrank, clung, swayed, shivered*). These collocates contrast with those of *he*, which include words such as *hummed* and *laughs.* Moreover, we can see from the collocates of *she* and *he* that it is *she* who is *choked* or *strangled* by a male criminal while it is *he* who *killed* a female. This paper discusses in great detail how Christie describes female in her works from the perspective of collocates of *she*.

Bibliography

Breiman, L. (2001). Random forests. *Machine Learning*, 45: pp.5-23.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Sinclair, J. and Jones, S. (1974). English lexical collocations: A study in computational linguistics. In Teubert, W. and Krishnamurthy, R. (eds.) (2007). *Corpus Linguistics: Critical Concepts in Linguistics. Vol.1*(pp. 223-269). London: Routledge.

Tsuchimura, N. (2016, September). *Stylistic Analysis of Agatha Christie's Works: Comparing with Dorothy Sayers*. Paper presented at the sixth annual conference of Japanese Association for Digital Humanities, Tokyo, Japan.

nodegoat: Enabling Explorative Research

Pim van Bree
pim@lab1100.com
LAB1100, The Netherlands

Geert Kessels
geert@lab1100.com
LAB1100, The Netherlands

Introduction

nodegoat allows scholars to build datasets based on their own data model and offers relational modes of analysis with spatial and chronological forms of contextualisation. By combining these elements within one environment, scholars are able to instantly process, analyse and visualise complex datasets relationally, diachronically and spatially; trailblazing.

nodegoat follows an object-oriented approach throughout its core functionalities. Borrowing from actor-network theory this means that people, events, artefacts, and sources are treated as equal: objects, and hierarchy depends solely on the composition of the network: relations. This object-oriented approach advocates the self-identification of individual objects and maps the correlation of objects within the collective.

Research Environment

nodegoat is a web-based research environment that facilitates an object-oriented form of data management with an integrated support for diachronic and spatial modes of analysis. This research environment has been developed to allow scholars to design custom relational database models. nodegoat dynamically combines functionalities of a database management system (e.g. Access/FileMaker) with visualisation possibilities (e.g. Gephi/Palladio) and extends these functionalities (e.g. in-text referencing, LOD-module) in one web-based GUI. As a result, nodegoat offers researchers an environment that seamlessly combines data management functionalities with the ability to analyse and visualise data.

The explorative nature of nodegoat allows researchers to trailblaze through data; instead of working with static 'pushes' – or exports – of data, data is dynamically 'pulled' within its context each time a query is fired. The environment can be used in self-defined collaborative configurations with varying clearance levels for different groups of users.

As a result of nodegoat's object-oriented set-up, everything is an object. In the case of a research project on correspondence networks, this means that a researcher would define three types of objects in nodegoat: 'letter', 'person', 'city'. Each object relates to an other object via relations (e.g. a letter relates to persons to identify the sender/receiver and this letters has been sent from/received in a city). In an extended research process, researchers could also define themselves as objects in the dataset, their sources or other datasets. Due to the focus on relations and associations between heterogeneous types of objects, the platform is equipped to perform analyses spanning multitudes of objects. By enriching objects with chronological and geospatial attributed associations, the establishment and the evolution of networks of objects is fully contextualised. In nodegoat, these contexts and sets of networked data can be instantly visualised through time and space.

This open-ended approach makes nodegoat different from tools like the Social Networks and Archival Context Project, Alan Liu's Research Oriented Social Environment, the Software Environment for the Advancement of Scholarly Research, Prosop, or tools with a main focus on coding of qualitative data as seen in various computer-assisted qualitative data analysis software. With its object-oriented approach, nodegoat facilitates the aggregation of collections, coding of texts, and analysis of networks, but models these methods towards the creation and contextualisation of single objects that move through time and space.

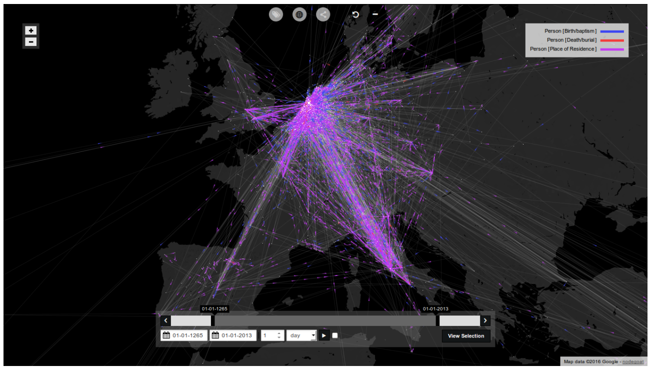
Facts & Figures

nodegoat is conceptualised and built by the independent research firm LAB1100, based in The Hague, The Netherlands. In order to share the functionalities of nodegoat with the scholarly community, scholars and research institutes are invited to use nodegoat for their own research purposes. Over 300 scholars have a personal research domain on nodegoat.net. Over 15 institutional partnerships have been established with universities, research institutes, and museums in The Netherlands, Belgium, Luxembourg, and Germany.

A nodegoat user forum and FAQ is hosted on the Historical Network Research website. In the course of 2018 an open source package of nodegoat will be released within the wider framework of the nodegoat community.

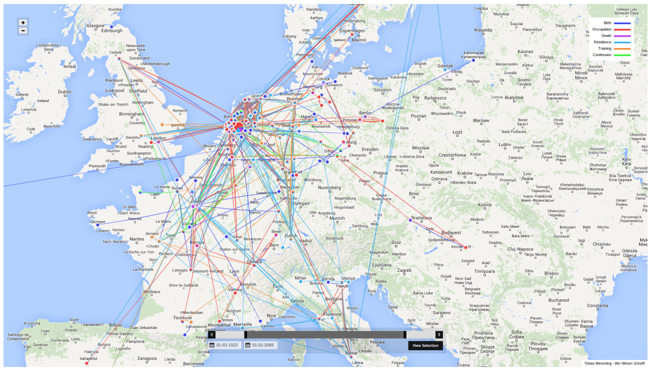
Examples of projects in nodegoat

Project 'Mapping Notes and Nodes in Networks' in collaboration with Huygens ING, University of Amsterdam, & KNIR



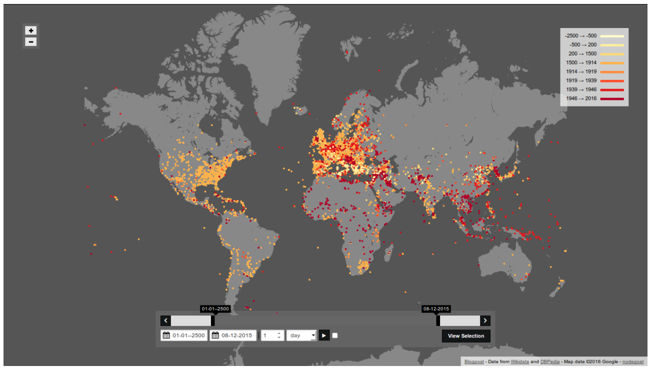
The whereabouts of over 20.000 people visualised through time and space in nodegoat http://mnn.nodegoat.net/viewer.p/1/47/scenario/17/geo/

Illustration of a personal research dataset in nodegoat



Geographical network visualisation in nodegoat by Tobias Winnerling for the project 'Wer Wissen Schafft'

A Wikidata/DBpedia Geography of Violence



Over 12.000 battles as described by Wikidata and DBPedia users visualised in nodegoat, http://nodegoat.net/blog.s/14/a-wikidatadbpedia-geography-of-violence.

Digital Humanities at Berkeley and the Digital Life Project

Courtney von Vacano
cvacano@berkeley.edu
UC Berkeley, United States of America

Abigail T. De Kosnik
adekosnik@berkeley.edu
UC Berkeley, United States of America

Stephen Best
sbest@berkeley.edu
UC Berkeley, United States of America

"The Digital Life Project" at UC Berkeley aims to reframe how we define digital humanities by bringing critical analysis to every stage of a highly collaborative and distributed research process. Below, we provide three examples of how this work is evolving at UC Berkeley.

To begin, incidents in the #BlackLivesMatter social movement raised our collective awareness and compelled us to rethink the digital humanities landscape on campus. We realized that understandings of race were being shaped by the dissemination of footage, images, and words. In particular, the dissemination of images of racial violence through media has begun to change the nature of questions in scholarship on race. Questions can no longer overlook a focus on the structure of those sources of data. Furthermore, in this rapid shift, social media must be understood as an infrastructure of hybrid online + offline human existence (i.e., the digital life).

We might perceive the dystopia of digital life in the preponderance of disturbing images of brutality and oppression, an archive that many—including victims of that violence—wish to reclaim. Conversely, we might understand the utopia of digital life to entail the affordance for social media users to document and share their experiences, instantaneously share information, and recruit attention, assistance, and aid to individuals and groups in need. In focusing on the broad rubric of "digital life," moreover, we want to support both dialogue in the broader university community and plans for focused research on a range of issues.

The first example of research within The Digital Life Project is a collaborative endeavor that is embedded in a digital humanities, ethnic studies methods course. Professor Keith Feldman worked closely with both the Ethnic Studies Library and the D-Lab to implement this project that included 65 undergraduate students, two graduate students, and a postdoctoral fellow, in addition to library staff. Students worked with digitized field recordings of events at Berkeley from the late 1960s from the Yuen Archive. The structuring and coding of the audio data went back to (a) all of the students for their individual research papers; (b) the Ethnic Studies Library to help seed a Thesaurus for the Yuen archive; and (c) the public in the form of more user-friendly audio files housed on the Internet Archive. The project impacted multiple constituencies in both immediate and long-term ways. By the end of this course, students were able to historicize the emergence and transformation of the humanities as a field of knowledge, use computational methodologies to conduct research on race and ethnicity, practice various ways to read closely, critically, and against the grain, while at the same time being exposed to the affordances and challenges of digital humanities.

Next, we embarked on a groundbreaking initiative to track and monitor types of Internet-based hate: The Online Hate Index (OHI). The goal is to create a comparative analysis of the online locations, sources, and relationships, and motivations behind online hate targeting different populations. Because of the marked rise of xenophobia in America today, the first installment focused on anti-Latino immigrant sentiments; subsequent studies will examine different forms of targeted hate. The D-Lab provided guidance on designing and implementing a computational research methodology that employed Machine Learning. The League of United Latin American Citizens (LULAC) provided specific information relating to anti-immigrant, anti-refugee, and anti-migrant populations and enabled the project findings to become actionable. In the future, ADL's Center for Extremism and Center for Technology and Society will provide expertise on the online locations and characteristics of hate groups, in-house computing resources, and the techniques used to terrorize groups.

After one year of the Digital Life Project, we reflect upon the lessons learned in generating a new body of research through a highly distributed infrastructure that optimizes (a) the knowledge of graduate students and staff consultants and (b) mentorship within working groups and research teams comprised of undergraduates, graduate students, postdocs, faculty, and staff. At the same time, we are keeping our sights on mobilizing activism in order to foment an activist/researcher stance. Our collaboratory research work clearly defines roles and tasks within the research teams and thereby provides much needed apprenticeship opportunities that underrepresented minorities lack.

Bibliography

Bailey, M. Z. (2011). All the Digital Humanists Are White, All the Nerds Are Men, but Some of Us Are Brave. http://journalofdigitalhumanities.org/1%E2%80%931/all-the-digital-humanists-are-white-all-the-nerds-are-men-but-some-of-us-are-brave-by-moya-z-bailey/ (accessed 6 April 2017).

Earhart, A. E. (2012). Can Information Be Unfettered?: Race and the New Digital Humanities Canon.

http://dhdebates.gc.cuny.edu/debates/text/16 (accessed 6 April 2017).

Gallon, K. (2016). Making a Case for the Black Digital Humanities. http://dhdebates.gc.cuny.edu/debates/text/55 (accessed 6 April 2017).

Liu, A. (2012). Where Is Cultural Criticism in the Digital Humanities? http://dhdebates.gc.cuny.edu/debates/text/20 (accessed 6 April 2017).

Neal, M. A. (2012, September 17). Race and the Digital Humanities, Left of Black (webcast), season 3, episode 1, John Hope Franklin Center. https://www.youtube.com/watch?v=AQth5_-QNj0 (accessed 6 April 2017).

Nelson, A. (2002, Summer). Introduction: Future Texts, Social Text, 20(2): 1–15.

Posner, M. What's Next: The Radical, Unrealized Potential of Digital Humanities. http://dhdebates.gc.cuny.edu/debates/text/54 (accessed 6 April 2017).

Virtual Hamlet: Combining Motion Capture and Real Time Digital Puppetry

Augustus Wendell
wendell@njit.edu
New Jersey Institute of Technology
United States of America

Louis Wells
wells@njit.edu
New Jersey Institute of Technology
United States of America

As a theatrical production, Hamlet poses a particular directorial challenge: the Ghost. How should an apparition be presented onstage alongside mortal characters? Traditional Shakespearean theatre presented apparitions using theatrical effects to visually separate these characters from the mortal characters. Modern theatrical productions must address this same issue with every new staging. Solving this question can help lay a clear conceptual footing for the rest of the production's conceptual considerations. A recent production in New York City in 2015 directed by Austin Pendleton opted to make the Ghost invisible to everyone but those guards who first encounter him, and once Hamlet sees him, only Hamlet. This decision rendered Hamlet mentally disturbed in the eyes of the audience (Isherwood 2017). Kenneth Branagh's cinematic Hamlet in 1996 created a ghost that had a "special effects and a horror-film look." (Maslin 1996) This approach favors brutal realism expected in modern cinema. These two approaches represent extremes in production solutions to the ghost; psychological illness, and pale faces in armor.

When Hamlet was first produced, the Elizabethan stage was full of special effects (Brockett, Franklin 2008). Shakespeare was a practitioner of special effects technology such as pyrotechnics, rope and pulley and trap doors. His commitment to authenticity and effect even led to the accidental destruction of a theater by fire in 1613 after the use of a cannon-based special effect. As conceived by faculty in a theatre program and a digital effects design department at the same university, the challenge of representing the ghost and the Shakespearean spirit of theatrical effect became an opportunity for collaboration. This was our opportunity to create a new type of ghost, one that bridges the traditions of theater and cutting edge interactive media.

Theater has a history of stage projections dating from the 1700's use of magic lantern devices (Figure 1) to the 1920's when innovators like Erwin Piscator began their experiments with the medium (Figure 2).

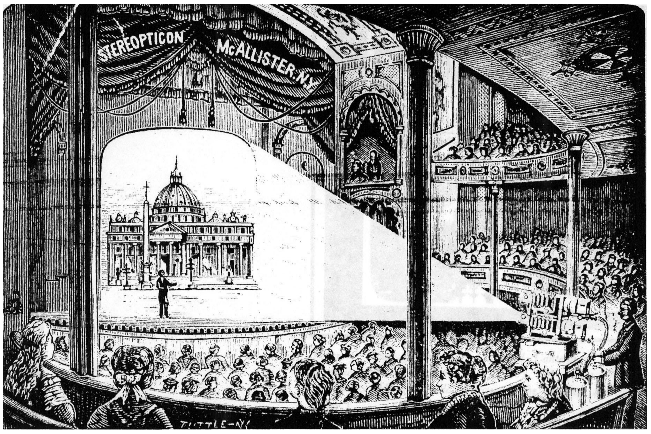


Figure 1. A magic lantern projector device.

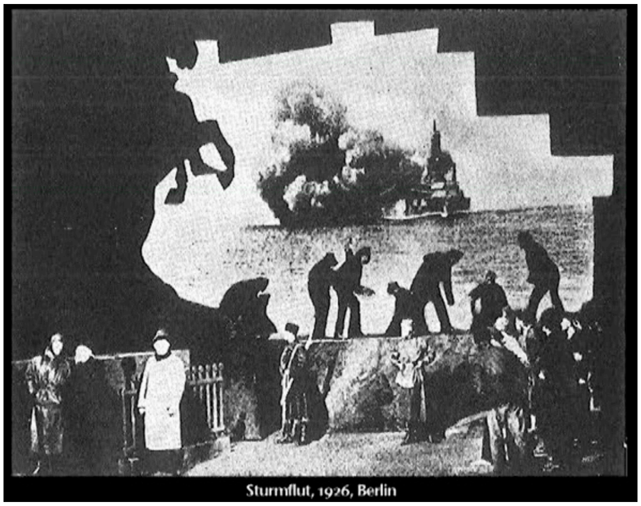


Figure 2. Erwin Piscator staging of *Sturmflut*, 1926 with projected imagery.

Throughout the twentieth century there was an extraordinary growth in the implementation of projections from experimental to mainstream theater (Theatre Communications Group 2011). While the size and complexity of projections has grown over time they have been until recently static or pre-recorded. With emerging technologies such as Microsoft's Kinect Camera and Derivative Software's TouchDesigner projections can be an interactive, responsive addition to the live theatrical experience. Our collaboration utilized these technologies in creation of our ghost.

Theatre has a long tradition of extending the body in performance. Costumes, makeup and prosthetics are the bread-and-butter of the art. Greek theatre used large scale masks and costumes to amplify theatrical expression, while Elizabethan theatre utilized extreme makeup applications and costuming for supernatural characters. Physical puppetry provides a clear mechanism to extend the performative and expressive capabilities of the actor. There are, however, some practical limitations to the physical puppetry approach: the movement and visual effect of the puppet is limited to the physical realities of the performer. Pushing beyond these boundaries, cinema has used digital motion capture to create such characters as Gollum in the Lord of the Rings trilogy and the recent Planet of the Apes films. Andy Serkis, an actor known for his motion captured performances, states, "Performance capture is a tool that allows actors to transform themselves into many different characters. You're not confined by physicality. You can play anything" (Hart 2017). Advances in virtual staging and performance capture, notably used in the production of the film Avatar, have extended motion capture for characters into the real time realm.

Although hybrid physical/digital stagings have been undertaken before (Meador, W. Scott, et al 2009, Bermudez et al 2002), the field has not kept pace with rapidly evolving technology. Our production combined recent advances in inexpensive real-time motion capture with the theoretical underpinnings of theatrical puppetry to perform a spatio-digital character in a live theatrical venue. A Kinect sensor on stage captured the movements of a physical actor performing the ghost. This data was manipulated in Derivative Software TouchDesigner and optically projected as an abstract digital apparition back onto the stage concurrent to the performance (Figures 3, 4). The virtual puppet became at once puppet and avatar, both extending and replacing the physical body of the actor. The dual space of this performance, half body-sized space of the actor and half virtualized space of the projected apparition, plays to the notion of the Cybrid space (Anders 1999).

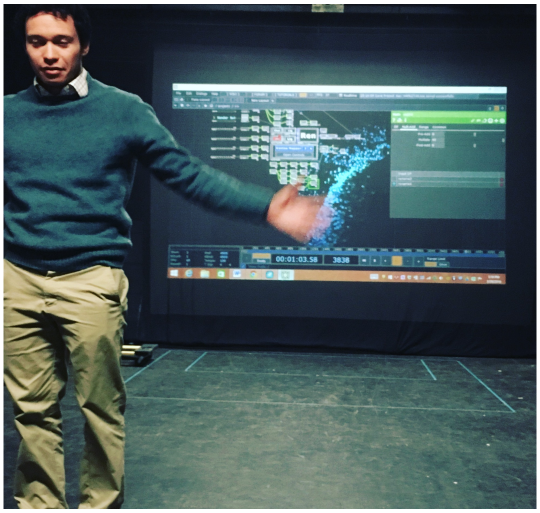


Figure 3. Technical rehearsal still from the NJIT 2016 Hamlet Performance.

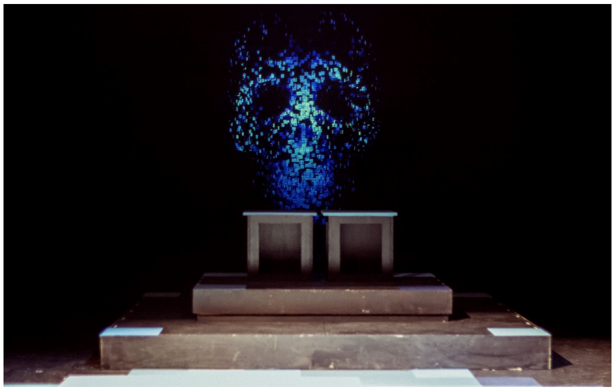


Figure 4. Production still from the NJIT 2016 Hamlet Performance.

This poster contextualizes, describes and presents the Spring 2016 theatrical production of Hamlet at the New Jersey Institute of Technology featuring a digitally created projected parametric Ghost character performed in real time by a motion-captured onstage actor.

Bibliography

Anders, P. (1999) *Envisioning Cyberspace*. 1st ed. New York: McGraw-Hill, 1999. Print.

Bermudez, J, Agutter, J, Syroid, N, Lilly, B, Sharir, Y, Lopez, T, Westenskow, D, and Foresti, S (2002). "*Interfacing Virtual and Physical Spaces through the Body: the cyberPRINT Project.*" In *Thresholds - Design, Research, Education and Practice in the Space Between the Physical and the Virtual: Proceedings of the 2002 Annual Conference of the Association for Computer Aided Design in Architecture*, 395-400. ACADIA. Pomona, California: Cal Poly, Pomona, 2002.

Brockett, O. G, and Hild, F. J. (2008). *History of the Theatre*. Boston: Pearson. Print.

Delbridge, M., and Tompkins, J. (2012). "Reproduction, mediation, and experience: virtual reality, motion capture and early modern theatre." *Space–Event-Agency–Experience*.

Hart, H. (2017). "When Will A Motion-Capture Actor Win An Oscar?". *WIRED*. N.p., Web. 7 Apr. 2017.

Isherwood, C. (2017). "Review: 'Hamlet' As An After-Party That Got Out Of Hand". *Nytimes.com*. N.p., Web. 1 Apr. 2017.

Maslin, J. (1996) "Hamlet". *Partners.nytimes.com*. N.p., 1996. Web. 1 Apr. 2017.

Meador, W. S., et al. (2004) "Mixing dance realities: collaborative development of live-motion capture in a performing arts environment." *Computers in Entertainment (CIE)* 2.2: 12-12.

Theatre Communications Group (2011) *University Library Home Page*, 2016 http://www.tcg.org/publications/at/dec11/projection.cfm. [31 March 2016].

Playing with Time

Jeri Wieringa
jwiering@gmu.edu
George Mason University, United States of America

The question of change over time is central for historians as they trace the contours of people, communities, organizations, and states. But computationally tracing changes in discourse over time is not a straightforward process with our current algorithmic tools and methods. While charting topics over time can give a general picture of the prevalence of ideas or discourses at various points, Benjamin Schmidt and others have raised concerns regarding the ways existing tools and methods model change over time. Most text-mining algorithms assume that time is linear and progressive, is experienced as such, and that the data is generally consistent over time. As Schmidt notes, even models such as Topics over Time, which was developed to account for shifts in language, are problematic because of the assumptions they make regarding how language changes (Schmidt 2012, see also Underwood, 2012, and Nowviskie, 2016).

If we allow that time is relative and that the experience of time shapes the ways a community and its discourses develop, we are faced with the challenge of how to incorporate varied experiences of time into our computational models. This poster will present my preliminary results applying various periodization schemes to track the development of Seventh-day Adventist discourses around salvation and health over the first 70 years of the denomination's existence.

Seventh-day Adventism is an apocalyptic and millennialist belief system, in that followers anticipate the imminent return of Jesus Christ and the corresponding end of the world. Born out of William Miller's teaching that 1843 (later 1844) would be the date of Christ's return, early Seventh-day Adventists reinterpreted the date in the wake of the Great Disappointment (the period after October 22, 1844, the date the Millerites believed Jesus would return. to signify the start of a new, final, and assumedly short phase in the work of salvation). They also adopted Sabbatarianism, holding Saturday, rather than Sunday, as the proper day for Christian observance. As such, early adherents operated within a changing temporal imaginary, organizing their weeks and years in contrast to their religious neighbors and anticipating, with varying degrees of urgency at different points in their history, the second coming.

The changing constructions of time within Seventh-day Adventism provide an instructive case study for examining how alternative structures and experiences of time might be modeled computationally. With a corpus of approximately 13,000 periodical issues split into nearly 200,000 pages, I am using Latent Dirichlet Allocation as implemented in Gensim to cluster the pages according to five different periodization schemes: no periodization (the whole corpus, spanning 70 years); periodization by decade (the corpus split by decade increments); cumulative periodization by decade (the corpus split by decade increments, but with each subsequent period added to the previous - i.e., 1844-1850, 1844-1860, 1844-1870, etc.); historical periodization (the corpus split according to "crisis" points in the denomination's history); cumulative historical periodization (the corpus split according to the historical periodization scheme above but with each subsequent period added to the previous).

Comparing these different schemes will enable me to explore whether and how periodization influences which discursive patterns I am able to surface computationally. To compare between the resulting models, I will evaluate on the following aspects: how well the documents are described by the topic labels assigned; the coherence of the topics, both internally and in relation to each other; and the visibility of change and development of topics over time.

My hypothesis is that the cumulative historical periodization will provide the richest picture of the shifts in the community's discourse. I anticipate that this will be visible through the percentage of documents labelled as relating to different topics, the appearance of new related topics over time, and different topic compositions at different points in time. However, I may conclude that the different periodization schemes provide little additional information. I propose that a negative conclusion is also important for the field and either result will provide productive information for developing computational methods that address the complexities that are at the heart of the humanities.

Bibliography

Schmidt, B. M. (2012) "Words Alone: Dismantling Topic Models in the Humanities" in *Journal of Digital Humanities* 2.1 http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ (accessed 01 Nov. 2016).

Underwood, T. (2012). "What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?" in *Journal of Digital Humanities* 2.1 http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/ (accessed 01 Nov. 2016)

Nowviskie, B. (2016) recent discussions of the logics of time and archival practices in "Speculative Collections," Nowviskie.org, 27 October 2016. http://nowviskie.org/2016/speculative-collections/ (accessed 01 Nov. 2016).

Digital Humanities Pedagogy in the University of Washington Information School MLIS Program

Helene C. Williams
helenew@uw.edu
University of Washington, United States of America

Digital Humanities Librarianship is a 3-credit elective open to University of Washington Masters of Library and Information Science students and a core elective in the Graduate Certificate in Textual and Digital Studies program. The course focuses on creating full contributors in the DH realm: a plethora (in academic terms, at least) of job postings for DH or Digital Scholarship librarians shows rapid growth in this area and a need for well-prepared individuals. No matter the physical location of DH centers—from libraries to academic departments to research centers—DH librarians are ideally suited for navigating the intellectual and geographic spaces of ideas, resources, and tools.

In addition to its unique hybrid format (synchronous residential and online), the course is also notable for providing a balanced approach to the theory vs. application tension typical of MLIS programs. Most DH courses in LIS departments focus on technologies, without concomitant attention to a foundational understanding of the varied research methods and resource usage patterns of humanities scholars. The DH Librarianship course aims to provide students with an understanding of how humanities scholars work, both traditionally and digitally, as well as familiarity with resources and tools used in digital humanities scholarship. The course also covers political and practical issues: what roles do librarians play in DH research, and what roles are situationally appropriate—tech guru, data cleaner, resource purchaser, equal collaborator? During the quarter, the course tackles questions of sustainability, accessibility, ethics, and equity in representation. Guest speakers include DH librarians (both with and sans MLIS), new DH faculty in various disciplines, and seasoned humanities researchers.

Assignments include disciplinary exploration, which allows students to explore resources and DH projects in philosophy, religion, fine and performing arts, languages and literatures. Student groups examine DH tools that range from timelines and mapping to text mining, information visualization, data cleaning, and network analysis, and create presentations from a shared corpus. In the final “DH consultation” assignment, students locate a project—in-process, abandoned, or “complete”—and propose options for library-based support, based on a disciplinary needs assessment. They also provide suggestions for strengthening the content, usefulness, or reach of the project, as well as a tools/usability assessment, which may include creating a prototype of an improved project.

The pedagogy reflects the multi-faceted disciplinary grounding and technological approach of the course content. My experience as an academic humanities librarian is bolstered by the research of many, including Melissa Terras, who discusses the need to identify core values and “hidden histories” of disciplines (2006), and Marcia Bates, who defines distinctions between disciplines such as the humanities and meta-disciplines like LIS (1999). In LIS we analyze the processes and domains of disciplines—in this case those involved in DH—and how those are represented, accessed, and given meaning across the corpora of recorded information. Mike Caulfield’s writing (2016) on multiple digital literacies also resonates with the pedagogies used, demonstrating the need for domain-grounded literacy to help students ascertain next steps and appropriate tools in their work with humanities scholars and projects.

The hybrid format of the class contributes to the collaborative pedagogy: local online and residential students attend the technology-enabled classroom in person, and others attend via the online Zoom classroom. Students can participate via audio or text, share screens, display presentations, or work in groups. Cameras in the classroom broadcast what is happening locally, and those attending online enable their webcams to be more fully present. In addition, several students each week attend via Kubi robots, which are iPads on movable stands that online users rotate to focus on different classroom activities. Guest speakers may attend in person, or they may present and engage with students via online options. These diverse modes of learning increase the students’ comfort with multiple technologies, which they are then more likely to use in their own research and teaching.

Learning outcomes include familiarity with the structure of knowledge in the humanities disciplines as well as the wide array of resources that provide reference and bibliographic support for research. Students are able to connect issues and concepts in DH to ongoing projects and scholarship, and they are able to articulate the ways in which library support fits into a changing paradigm of research in the humanities disciplines. In addition, students understand issues concerning equity, representation, and emotional labor in conjunction with digital humanities librarianship.

This poster will highlight the multiple teaching and learning techniques used in the course and student projects. It will also showcase the physical and virtual teaching and learning spaces.

Bibliography

Bates, M. J. (1999). “The Invisible Substrate of Information Science.” *Journal of the American Society for Information Science*, 50(12): 1043-50.

Caulfield, M. (2016). “Yes, Digital Literacy. But Which One?” *Hapgood.* December 19. Retrieved from https://hapgood.us/2016/12/19/yes-digital-literacy-but-which-one/

Terras, M. (2006). "Disciplined: Using Educational Studies to Analyse ‘Humanities Computing'." *Literary and Linguistic Computing*, 21: 229-46.

Effective Identification of Citations in the Kanseki Repository

Christian Wittern
cwittern@gmail.com
Kyoto University, Japan

Introduction

The Kanseki Repository is a large repository of premodern Chinese texts. Currently it holds more than 9000 texts, covering all periods of Chinese history from early antiquity to the beginning of the 20th century. The repository is organized into 6 top-level categories and offers full-text search across all textual variants.

Since its opening to the public in March 2016, a frequent request from users was to be able to find texts related to a certain text, especially to investigate and evaluate textual dependencies. This poster reports on some experiments to find an effective way to respond to this requests.

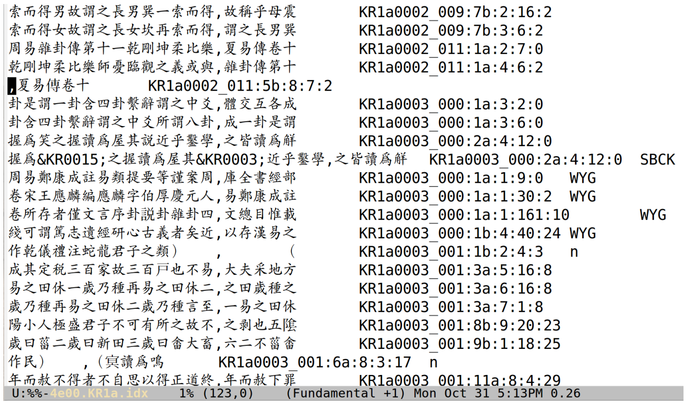
What is needed to solve this problem is an efficient way to identify text passages that are derived from other, earlier texts (based on the assumption that the texts in question can in fact be reliably dated). Such passages will be called **citations** here, although the usage here is not limited to true citations, but also includes quotations from memory, paraphrases and allusions – cases where the reference does not follow the exact wording of the source. As an additional complication, we need to take into account the possibility that the received text differs from the text available to the author of the text that contains the reference.

Related work has examined plagiarism detection (Gipp and Meuschke, 2011 and Schultz, 1999), but the approach taken here makes direct use of some of the unique features of the repository and the index built for it and seems thus to be more efficient than general purpose algorithms. Admittedly, this has not been verified empirically and thus may be reasonably rejected as not relevant. However, the purpose of this presentation is not to compare algorithms and their efficiency when applied to the material here, but rather to collect some low-hanging fruit that became available due to the way the full-text index is constructed.

Identification of Citations

Index

Since a complete index has already been built for the full-text search, all experiments make use of this index (Mandoku 2016). The index is constructed by moving an n-gram window over the text and saving entries at appropriate locations. The resulting raw index is then read by a grep-like program to generate the display. The search display is designed to show the **keyword in context** (KWIC) so some characters are needed in front of the match character; these are appended after a comma character in the index entry. We built the index with a window between 10 and 25, since larger indexes considerable increase the required space and smaller builds will have too little information in the KWIC display. The index also contains information to identify the text, the location of the index excerpt and some information about the context of the match. Figure 1 shows a typical example of such an index for a 21-gram window.



```
索而得男故謂之長男巽一索而得,故稱乎母震        KR1a0002_009:7b:2:16:2
索而得女故謂之長女坎再索而得,謂之長男巽        KR1a0002_009:7b:3:6:2
周易雜卦傳第十一乾剛坤柔比樂,夏易傳卷十        KR1a0002_011:1a:2:7:0
乾剛坤柔比樂師憂臨觀之義或與,雜卦傳第十        KR1a0002_011:1a:4:6:2
,夏易傳卷十        KR1a0002_011:5b:8:7:2
卦是謂一卦含四卦繫辭謂之中爻,體交互各成        KR1a0003_000:1a:3:2:0
卦含四卦繫辭謂之中爻所謂八卦,成一卦是謂        KR1a0003_000:1a:3:6:0
握爲笑之握讀爲屋其説近乎鑿學,之皆讀爲觧        KR1a0003_000:2a:4:12:0
握爲&KR0015;之握讀爲屋其&KR0003;近乎鑿學,之皆讀爲觧   KR1a0003_000:2a:4:12:0  SBCK
周易鄭康成註易類提要等謹案周,庫全書經部        KR1a0003_000:1a:1:9:0   WYG
卷宋王應麟編應麟字伯厚慶元人,易鄭康成註        KR1a0003_000:1a:1:30:2  WYG
卷所存者僅文言序卦説卦雜卦四,文總目惟載        KR1a0003_000:1a:1:161:10        WYG
綫可謂篤志遺經研心古義者矣近,以存漢易之        KR1a0003_000:1b:4:40:24 WYG
作乾儀禮注蛇龍君子之類）    ,        （        KR1a0003_001:1b:2:4:3   n
成其定税三百家故三百戸也不易,大夫采地方        KR1a0003_001:3a:5:16:8
易之田休一歲乃種再易之田休二,之田歲種之        KR1a0003_001:3a:6:16:8
歲乃種再易之田休二歲乃種言至,一易之田休        KR1a0003_001:3a:7:1:8
陽小人極盛君子不可有所之故不,之剥也五陰        KR1a0003_001:8b:9:20:23
歲曰菑二歲曰新田三歲曰畬大畜,六二不菑畬        KR1a0003_001:9b:1:18:25
作民）    ,（寘讀爲鳴        KR1a0003_001:6a:8:3:17  n
年而赦不得者不自思以得正道終,年而赦下罪        KR1a0003_001:11a:8:4:29
 U:%%-4e00.KR1a.idx    1% (123,0)    (Fundamental +1) Mon Oct 31 5:13PM 0.26
```

Figure 1. Effective Identification of Citations in the Kanseki Repository

Algorithm

To find citations in the indexed files, a window of the same size as the index window is moved over the text passage under investigation. A search is initiated for a string of **n** characters, starting at the first index position. In the example in Figure 1 this would be starting at position 6, since there are 5 characters after the ","; these characters are preceding the indexed characters in the text and have therefore been moved to the back. A query expansion is used for this search, in order to catch character variants in this initial selection of index lines. A large value of **n** will increase the probability that citations are not found due to slight positional variations in the text, while a small value of **n**, such as 1, will select many lines that are not relevant and will thus increase the processing times. Experiments have shown that a value of 2 or 3 for **n** gives a good optimum for precision and recall. Positional values are also registered to better demarcate the citation boundaries.

The selected lines have to be post-processed to restore the original sequence as found in the text. The line will then be compared to the window of the text passage, with scores given for each match; high scores are taken to be a citation and are marked for further processing. The best results so far have been achieved for a cumulated score of n-gram matches for values of **n** from 1 to 3, but additional experiments are planned. Conclusive results will be reported in the poster presentation.

Additional expected results

With the method introduced here, it becomes feasible to investigate potential citations for whole texts. We plan to build a heat-map of a text with passages that have been cited coloured according to their frequency. This will enable new ways of exploring the intertextuality of texts and will provide new evidence for the history of ideas and flow of intellectual debt in the history of Chinese thought. For the presentation of the poster, we show a preliminary investigation of some key texts of Chinese philosophy as a proof of concept.

We also hope to identify a set of key phrases and look at their usages over time, and in different schools of thought, to see what kind of trends can be seen there.

Bibliography

Gipp, B. and Meuschke, M. (2011). "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence". In: *Proceedings of the 11th ACM symposium on Document engineering* (DocEng '11), Mountain, View, CA, USA, 2011. doi: 10.1145/2034691.2034741.

Mandoku (2016). Mandoku project (Source code) https://github.com/mandoku/mandoku (accessed 2016-10-31).

Schultz, R.L. (1999) *The search for quotation : verbal parallels in the prophets,* Sheffield, Sheffield Academic Press, 1999.

Wittern, C. (2014). "Kanripo and Mandoku: Tools for git-based distributed repositories for premodern Chinese texts", in *Digital Humanities 2014 Book of Abstracts*, 2014, p. 408-409.

Wittern, C. (2016). Special issue Kanseki Repository, *CIEAS Research Report* 2015, Kyoto 2016.

Collecting the Name of a Historical Person from Related Historical Material

Taizo Yamada
t_yamada@hi.u-tokyo.ac.jp
The University of Tokyo, Japan

Satoshi Inoue
inoue@hi.u-tokyo.ac.jp
The University of Tokyo, Japan

This proposal describes a method of collecting and managing the name of a person from historical materials related to Japanese history from the 8th century until the 19th century.

There are several tasks related to personal name such as collecting, identifying, representing. For historians and researchers, the recording and management of a personal name is an integral (and unavoidable) task. In Japan, historical materials are held by both a public institute and a private house in which they are found. Because there are no catalogs listing all of the material, the research staff at our institute (the Historiographical Institute of the University of Tokyo) spend several weeks each year investigating and examining historical records kept in Japan and abroad. The staff has over 100 years' combined experience conducting these examinations, and their journeys have taken them all over Japan as well as to many different parts of the globe. Ho (2015) and Bol et al (2015) introduced a method of personal name management in the context of Chinese historical materials (like a "地方志 (difangzhi)") that used the China Biographical Database (CBCB) as a biographical dictionary. In Japanese history, there is no such exhaustive encyclopedia or dictionary for a name of a historical person.

A personal name appears in a variety of historical materials such as an old diary, an old document, a classical book, a family pedigree, a document related to appointment, and so on. The materials contain diverse name representations, such as a real name or an original name, a nickname, an epithet, a role name which indicates a person, and a Kao (which is a stylized signature or a mark), among others.

We constructed a repository which can store a collection containing both the names and the historical materials in which they appear. There are several different representations of a name in each material type. In order to represent and preserve these variations, the repository can store a variety of name data regardless of differences in data schema. In order to ensure efficient data-searching, we prepared common metadata that consist of an identifier, a personal name, a correspondence of the name, a related resource, a database where the name has been stored if such a database has already been constructed and the date on which the name appeared in the resource.

The name data can be expressed as RDF data and searched with SPARQL in the repository which provides an RDF store and a SPARQL endpoint. Furthermore, the repository can integrate the name data regardless of differences among personal name metadata schema, and provide the result of the integration. In addition, we will show an application for representing the outcomes of search results.

Bibliography

Ho, H. I. (2015). MARKUS – A Fundamental Semi-automatic Markup Platform for Classical Chinese. Proceedings of the 2015 International Conference on Digital Humanities.

Bol, P., Liu, Ch.-L., and Wang, H. (2015). Mining and Discovering Biographical Information in Difangzhi with a Language-Model-based Approach. Proceedings of the 2015 International Conference on Digital Humanities.

Interpreting Racial Identities and Resistance to Segregation in the Digital Sphere

Mishio Yamanaka
yamanaka@live.unc.edu
UNC Chapel Hill, United States of America

This poster demonstrates my digital history project, “The Fillmore Boys School in 1877: Racial Integration, Creoles of Color and the End of Reconstruction in New Orleans,” (http://fillmoreschool.web.unc.edu) which visualizes how, in 1877, Creoles of color, a group of francophone Catholics of interracial descent, responded to the city school board’s racial segregation plan. Through geo-spatial analysis and virtual representation of information of the Fillmore School student register, the project examines Creole families’ resistance to segregation and facilitates public understanding of ambiguous racial identities that become often invisible in the black and white racial dichotomy of United States history.

The Fillmore School served as a catalyst for Creoles’ ideal of racial equality and the center of their resistance to segregation. During Reconstruction, Creoles of color led the legal, political and grassroots campaign that ultimately resulted in the desegregation of approximately one-third of the public schools in New Orleans between 1871 and 1877. Fillmore was one of the integrated schools that many Creoles of color attended until the school board designated it as white-only in 1877. Its 1877 register contains 658 students’ individual information, including name, residential address, age, admission date, and parents’ name and their occupation. While the register does not list race, it includes notations of transfer to a “colored” school for 16 students. In addition, my digital research and mapping revealed that at least forty Creole students requested admission to the school as a sign of protest. This project thus demonstrates a complex process of segregation, resistance and racial identities among Creoles of color in New Orleans.

The project has undergone four processes: 1) transcription, 2) data collection, 3) mapping, and 4) website building. First, I digitized the entire 1877 Fillmore School student data. Second, I searched the 1870 and 1880 censuses and city directory information of the students by using ancestry.com. Next, I deployed ArcGIS and georeferenced the 1883 Robinson Atlas, the most comprehensive map of New Orleans for the era, and collected students’ latitude-longitude information. I used ArcGIS because it allowed me to publish the data as an interactive map while simultaneously conducting geospatial analysis. The project processed 566 student data entries to be published as an interactive map to the public. Finally, I built a website to contextualize the map.

The main part of my poster demonstrates the Arc-GIS map. It pays particular attention to the ways in which the map shows students’ residences with their variable racial information taken from the register, the 1870 and 1880 censuses and city directories. The map displays the complex reality of segregation in New Orleans. The censuses classified many Creole students as mulatto or white. However, the census racial information does not always correlate with that of the school register. The poster also focuses on how the map represents multimedia information about Creoles of color and shows their admission attempts as a tactic to resist racial resegregation. I will also discuss challenges I faced in mapping, including the lack of complete data.

The poster also introduces my project website that offers a virtual space for users to consider responses among ordinary Creoles of color confronting segregation and understand the complicated history of race and ethnicities in the United States. The website currently offers Creole family narratives to further contextualize the map. It also provides a short history of the Fillmore School and school racial policies in New Orleans.

Last, this poster discusses the future plan for the project. First, I will deploy social network analysis of the Creole parents and children at Fillmore and examine how their close social relationships contributed to resistance. Second, I plan to examine the Fillmore School history in the twentieth century by using oral interviews to reflect another desegregation attempt during the civil rights movement in the 1960s. Overall, this poster argues for digital ways to illuminate ambiguous identities and a wide range of struggles against racism in the late nineteenth century United States and discusses how digital humanities enrich historical research.

Bibliography

Algeo, K., Epperson, A. and Blunt, M. (2011). "Historical GIS as a platform for public memory at Mammoth Cave National Park." International Journal of Applied Geospatial Research, 2(4): 19-37.

Devore, D. E. and Logsdon, J. (1991). Crescent City Schools: Public Education in New Orleans, 1841-1991. Lafayette: Center for Louisiana Studies, University of Southwestern Louisiana.

Gregory, I. N. (2003). A Place in History: a Guide to Using GIS in Historical Research. Oxford: Oxbow Books.

The Game of Writing: Gamification and Social Commenting in Writing Instruction

Jinman Zhang
jinman@ualberta.ca
University of Alberta, Canada

Roger Graves
graves1@ualberta.ca
University of Alberta, Canada

Heather Graves
hgraves@ualberta.ca
University of Alberta, Canada

Geoffrey Rockwell
grockwel@ualberta.ca
University of Alberta, Canada

How can we test gamification and social learning in online writing environments? This poster will demonstrate an online writing environment, GWrit (Game of Writing), where students can comment on each other's writing and where they get rewards for task activity in a cooperative environment (gamification in this instance is not competitive). The application of game-based learning strategies to teaching writing shows promise, with one study reporting that their role-playing game improved the quality of student writing (Wang, Chen, Chang & Chan, 2016). GWrit has been developed by a cross-disciplinary team of academics and programmers at the University of Alberta over the past three years and has been used with over 1000 students.

This poster/demonstration brings together research on GWrit from the following perspectives: gamification, social-network influenced peer review, and the task completion structures. Our research will be summarized on the poster and we will demonstrate the environment during the poster session.

Gamification. Deterding (2011) defined gamification as the use of game elements and game design techniques in non-game contexts to engage people in solving problems. Gamified environments employ a number of mechanisms to encourage people to engage with them (Dicheva et al., 2015), and existing studies prove that gamified learning environments create deeper engagement of students (Barata, Gama, Jorge & Gonçalves, 2013; Fitz-Walter, Tjondronegoro & Wyeth, 2012). Our gamified environment has many "surface" (award trigger systems, competitive environments, badges, and ranks) as well as "deep" gamification components (task completion structures, social commenting support, public posting of draft documents).

When demonstrating GWrit we will first introduce the way the system was designed to support experimenting with commenting and gamification, and then show the gamification rule editing environment we developed. Our working hypothesis was that users want information about what they are doing and that gamification can be a playful way of representing that information back to the users so that they can make decisions and possibly be motivated differently.

Social-network influenced peer review. Social networking has also been shown to have a positive influence on students' academic learning (Tian, Yu, Vogel & Kwok, 2011; Tsuiping , 2016). Within GWrit we provided an environment where students have the opportunity to both read and then comment on each other's drafts, a technique that has been shown to improve writing (Schunn, Godley, & DeMartino 2016; Ion, Barrera-Corominas & Tomàs-Folch, 2016). Reading skill is directly linked to writing improvement, particularly if students are reading texts similar to the texts they are trying to produce (Hansen, 2013). GWrit allows students to post drafts of their documents for review and comment by other students, peer tutors, and graders. The writer of the draft can respond to each comment, and they are also likely to reciprocate by reading and commenting on the drafts of the students who gave them comments. Micro-networks of comments sprout up within the comments on these texts. Because students in the writing course version of GWrit have the option of working on one of three different assignments in each module, larger, informal networks of students who are working on the same assignment also coalesce. The writing course version of GWrit has four main three-week long modules with a choice of three assignments in each module; the social networks re-form at the end of each module. Our early assessments of students and commenting in the writing course confirm what others have reported: that peer feedback is as valuable as instructor feedback (Guasch, Espasa, Alvarez, Kirschner 2013).

Task completion structures. Third, the poster will deal with the role task completion structures play in motivating learning. GWrit incorporates three task completion structures to help students: the course completion fuel gauge, a task list, and assignment deadlines. All three of these task completion structures were used in WRS 102 in winter 2016 term and the first and the last structures were used in fall 2015 term. The data on the task structure from a course survey will be shown and discussed in the poster.

In our discussion of the research on the poster we will summarize interview data on why we think the various aspects of the system work, which areas we think need to be improved to work better, and how we intend to transform a curriculum-based tool (the course-based version of GWrit) into a free-standing, web-based site.

Bibliography

Barata, G., Gama, S., Jorge, J., & Gonçalves, D. (2013). "Improving Participation and Learning with Gamification." In P*roceedings of the First International Conference on Gameful Design, Research, and Applications* (pp. 10–17). New York, NY, USA: ACM. http://doi.org/10.1145/2583008.2583010

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). "From Game Design Elements to Gamefulness: Defining 'Gamification'." In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). New York, NY, USA: ACM. http://doi.org/10.1145/2181037.2181040

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). "Gamification in Education: A Systematic Mapping Study." *Journal of Educational Technology & Society,* 18(3), 75–88.

Fitz-Walter, Z., Tjondronegoro, D., & Wyeth, P. (2012). "A Gamified Mobile Application for Engaging New Students at University Orientation." In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (pp. 138–141). New York, NY, USA: ACM. http://doi.org/10.1145/2414536.2414560

Guasch, T., Espasa, A., Alvarez, I.M., & Kirschner, P. A. (2013). "Effects of feedback on collaborative writing in an online learning environment." *Distance Education*, 34:3, 324-338, DOI: 10.1080/01587919.2013.835772

Ion, G., Barrera-Corominas, A., & Tomàs-Folch, M. (2016). "Written peer-feedback to enhance students' current and future learning." *International Journal of Educational Technology in Higher Education,* 13:15. DOI 10.1186/s41239-016-0017-y

Hasan, L., Morris, A., & Probets, S. (2009). "Using Google Analytics to Evaluate the Usability of E-Commerce Sites." In M. Kurosu (Ed.), *Human Centered Design* (pp. 697–706). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-02806-9_81

Hansen, J. (2013). "The Language Arts Interact." In Flood, J., Lapp, D., Squire, J.R., and Jensen, J. M. *Handbook of Research on Teaching the English Language Arts*. Mawah NJ: Erlbaun.

Schunn, C., Godley, A., & DeMartino, S. (2016). "The Reliability and Validity of Peer Review of Writing in High School AP English Classes." *Journal of Adolescent Literacy* 60:1, 13-23.

Tsuiping C. (2016). "Technology-supported peer feedback in ESL/EFL writing classes: a research synthesis." *Computer Assisted Language Learning,* 29:2, 365-397. DOI: 10.1080/09588221.2014.960942

Tian, S. W., Yu, A. Y., Vogel, D., & Kwok, R. C. W. (2011). "The impact of online social networking on learning: a social integration perspective." *International Journal of Networking and Virtual Organisations*, 8(3/4), 264. http://doi.org/10.1504/IJNVO.2011.039999

Wang, J. H., Chen, S. Y., Chang, B., & Chan, T.W. (2016). "From integrative to game-based integrative peer response: high ability versus low ability." *Journal of Computer-assisted Learning,* 32, 170–185. doi: 10.1111/jcal.12125

Pre-Conference Workshops and Tutorials

Construire sa bibliothèque numérique avec l'outil libre Omeka pour valoriser ses documents numérisés

Daniel Berthereau
daniel@berthereau.net
Indépendant, France et Canada

Description

Cet atelier d'initiation a pour objectif de permettre aux participants de construire leur première bibliothèque numérique libre et ouverte. Il vise à montrer que créer et utiliser une bibliothèque virtuelle est à la portée de tous, que l'on soit une grande institution, une bibliothèque, un service d'archives, un laboratoire ou une structure quelconque disposant de fonds documentaires ou archivistiques. C'est même possible d'en bas, par des bibliothécaires et par les chercheurs eux-mêmes, par exemple pour présenter les documents utilisés dans une monographie.

Il s'effectuera sur la base du logiciel libre Omeka S et sera divisé en quatre temps : réflexions autour de la conception de sa bibliothèque numérique, découverte et installation d'Omeka et de ses modules, création de documents et import de métadonnées et de fichiers, création d'une exposition virtuelle.

La bibliothèque numérique, un outil de base pour les humanités numériques

Une bibliothèque numérique est un site web qui permet de valoriser tous les documents numérisés dans des fichiers que l'on a dans ses armoires et ses tiroirs et qui dorment dans des disques durs, des CDs, dans des caves humides ou des greniers ouverts à tout vent, Les fichiers peuvent être issus de la numérisation de livres, de journaux, de cahiers, de registres, d'archives, de travaux étudiants, de cartes, de tableaux, de photographies, ou de toute autre image. Ce peut aussi être des fichiers audio ou vidéos ou d'autres formats issus de numérisations ou nativement numériques (pdf, images 3D de sculptures ou d'objets…).

De grandes bibliothèques numériques existent, comme Internet Archive, Hathi Trust, Europeana ou Gallica pour la Bibliothèque nationale de France, mais il en existe aussi des milliers d'autres pour des institutions plus petites ou pour des projets ou des fonds plus spécifiques. Historiquement, elles ont d'abord servi à publier des documents et à simplifier leur consultation sans contrainte de personnel et sans contrainte matérielle, notamment pour les documents fragiles ou précieux, avant que l'on rende compte que l'on pouvait faire beaucoup plus.

En effet, l'ouverture d'une bibliothèque numérique simplifie non seulement l'accès aux fonds, mais elle en démultiplie les possibles. Le simple fait qu'un fonds soit publié sur l'internet implique qu'il soit référencé automatiquement et qu'il soit donc mieux connu de la communauté des chercheurs et du public, qui pourront les découvrir, en prendre connaissance et les étudier facilement. Cela constitue une fondation pour la recherche dans les humanités numériques. Ensuite, comme dans une bibliothèque physique, dans un musée, dans des centres d'archives ou dans une galerie, les documents catalogués peuvent être valorisés par la création de collections et d'exposition temporaires ou permanentes. L'avantage de la dématérialisation est que les objets numériques peuvent appartenir à plusieurs collections et se trouver dans plusieurs salles ou sur plusieurs étagères, contrairement aux objets physiques. Quant aux expositions virtuelles, elles sont accessibles à tous et sans limitation, sans qu'il soit nécessaire de déplacer les documents ni que les chercheurs et les visiteurs ne se rendent physiquement sur place.

Au-delà de la publication et de la mise à disposition de documents, les bibliothèques numériques constituent aussi des outils de recherche. Les chercheurs les utilisent pour leurs propres recherches. En effet, la recherche est simplifiée en ce que la découverte, la comparaison et la confrontation des documents peut se faire en quelques clics. C'est d'autant plus vrai pour les documents textuels qui ont fait l'objet d'une océrisation ou d'une transcription, mais c'est aussi le cas pour tous les autres types de documents qui disposent de notices descriptives complètes et standardisées. Les chercheurs peuvent aussi les utiliser en aval pour publier leurs propres documents ou les documents libres de droits qu'ils ont utilisés, de la même manière qu'ils incluaient autrefois des photocopies dans leurs articles ou dans les annexes de monographies.

Il est aussi possible de faire appel à l'intelligence du public et à son temps libre dans une logique de science participative, par exemple pour identifier des informations sur des documents, pour compléter des notices, pour transcrire ou traduire des manuscrits ou simplement pour les étiqueter (ajouter des « *tags* ») dans une logique de folksonomie ou pour les commenter. C'est utile aussi bien pour les structures qui n'ont pas de moyens que pour celles qui n'ont pas de temps. Naturellement, cela implique un travail de communication qui dépasse la mise en place de l'outil.

Enfin, de par leur simple présence sur la toile, les documents et leurs métadonnées peuvent être échangés et réutilisés facilement par les autres sites et par les autres services de l'internet, dans la logique du web sémantique. On évoque ce dernier depuis longtemps, mais sa possibilité concrète et à la portée de tous est récente.

Le choix d'Omeka, un outil libre conçu pour le web sémantique

L'atelier d'initiation est conçu sur la base d'Omeka , un logiciel libre proposé par un centre de l'université George Mason en Virginie (États-Unis), le « Roy Rosenzweig Center for History and New Media », également créateur de Zotero, un logiciel de gestion bibliographique largement utilisé par les chercheurs. Omeka a été créé en 2007 et fête donc ses dix ans, avec une nouvelle version « Omeka S » spécialement dédiée au web sémantique.

La version actuelle, désormais renommée « Omeka Classic » et qui continue d'être maintenue, a pour objectif de reproduire le fonctionnement d'une bibliothèque et en reprend les concepts : création de notices de documents autour du Dublin Core ou de champs libres, création de types de documents, rassemblement des documents dans des collections, création d'expositions virtuelles, partage des données avec OAI-PMH, protocole d'échange largement utilisé dans les bibliothèques. Bien sûr, s'agissant d'un outil destiné à diffuser des documents souvent anciens, il n'y a pas de gestion des droits ni des prêts. Techniquement, Omeka est un CMS classique créé autour de php et mysql, basé sur la plateforme Zend, qui gère les utilisateurs, les vues, les urls, etc. Il dispose d'une api et il est extensible. Une dizaine de thèmes, avec un affichage réactif adapté à tous les écrans (« *responsive design* » / « *mobile first* »), et une centaine d'extensions (gestion des métadonnées, gestion des fichiers, visionneuses, contributions, diffusion) ont été développés par des universités, des agences web ou des prestataires, quasiment tous libres.

La nouvelle version « Omeka S », est une réécriture complète du logiciel et en reprend tous ces concepts, mais va au-delà en tirant pleinement parti des principes et des standards du web sémantique. Elle est en effet basée sur la spécification rdf, sur des ontologies (par défaut Dublin Core Terms, Bibliographic Ontology [bibo] et Friend of a Friend [foaf]) et sur le format json-ld. Elle est multisite (une seule installation d'Omeka permet de gérer plusieurs sites si besoin, avec des administrateurs indépendants). Son ergonomie a été repensée et les utilisateurs peuvent disposer de formulaires de saisie adaptés selon les types de documents. Elle est multilingue tant au niveau de l'interface qu'au niveau des métadonnées. Elle se concentre désormais sur la diffusion et des extensions permettant de synchroniser les documents issus d'archives institutionnelles telles que D-Space ou Fedora. Une dizaine d'extensions ont déjà été mises à niveau. Enfin, elle conserve la finalité d'Omeka Classic, celle d'être un outil simple ayant pour objectif effectif de mettre des concepts de la documentation numérique à disposition des non-spécialistes.

Finalité de l'atelier d'initiation

L'intérêt fondamental de cet atelier d'initiation pour la communauté des humanités numériques réside dans le concept même de bibliothèque numérique. Les bibliothèques et leur déclinaison numérique demeurent en effet un lieu essentiel pour les étudiants et les chercheurs, au cœur de la vie universitaire et de la recherche, quelle que soit la discipline. C'est aussi un lieu où les chercheurs peuvent présenter non seulement les résultats de leurs recherches, mais aussi les documents qu'ils ont utilisés pour celles-ci, lorsqu'ils sont libres de droits. Enfin, c'est un lieu où peuvent se retrouver et échanger les chercheurs et le grand public ou le public des amateurs éclairés, notamment dans le cadre de la recherche participative.

Présentateur

Daniel Berthereau

Daniel Berthereau, constructeur de bibliothèques numériques, a réalisé plusieurs bibliothèques virtuelles, depuis leur conception jusqu'à leur publication en passant par la préparation du cahier des charges, la normalisation des fichiers, la standardisation des métadonnées et leur import pour des bibliothèques et des centres de recherche de différentes disciplines. Il développe également des outils de conversion de métadonnées documentaires ainsi que des modules et des thèmes pour Omeka (cf. https://github.com/Daniel-KM).

Public visé et remarques techniques

Toute personne disposant de fonds numérisés et souhaitant les valoriser sur l'internet, en particulier les dépositaires de fonds, les bibliothécaires, les documentalistes chargés du système d'information, les curateurs, les archivistes et les chercheurs. Jusqu'à 20 personnes, avec leur ordinateur portable, une connexion internet et éventuellement quelques documents (fichiers et notices) pour expérimenter l'outil.

Programme

Cet atelier d'initiation de trois heures vise à présenter la création d'une bibliothèque numérique par sa mise en pratique réelle. Il s'agit d'une version synthétique d'une formation réalisée auprès de documentalistes et de chercheurs de l'Université Paris Ouest Nanterre (Labex « Des passés dans le présent »), auprès d'étudiants en master « Médiation culturelle et patrimoine numérique » de l'Université de Paris Vincennes-Saint-Denis et pour plusieurs bibliothécaires et conservateurs.

1. **Pourquoi construire une bibliothèque numérique avec Omeka ?**
    - Réflexions autour de la conception de sa bibliothèque numérique
    - Omeka et ses concurrents dans le cadre du web sémantique
    - Présentation d'Omeka (Omeka Classic et Omeka S) et de ses concepts

2. **Installation du logiciel**
    - Installation d'Omeka (ou utilisation d'instances préinstallées)
    - Présentation, choix et installation des modules
    - Choix et personnalisation du thème (interface graphique publique)

3. **Ajout des documents et des contenus dans le catalogue**
   - Création de son premier document numérique
   - Standardisation des métadonnées pour le web sémantique
   - Import de fichiers et de métadonnées en lot
4. **Valorisation des documents et des notices**
   - Modules de science participative
   - Modules d'exposition et d'échanges de données
   - Création d'une exposition virtuelle

Advancing Linked Open Data in the Humanities

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Abigel Lemak
alemak@uoguelph.ca
University of Guelph, Canada

Kim Martin
kmarti20@uoguelph.ca
University of Guelph, Canada

Robert Warren
rwarren@math.carleton.ca
Carleton University, Canada

Brief Description

Since its inception, Linked Open Data (LOD) has been primarily about publishing and defining data standards. As the technologies have matured and the amount of data available for consumption has dramatically increased, the question of consumption and processing is now at the forefront.

- What will scholars do with a large universe of linked open data?
- What initiatives are needed? Which tools do we need most?
- If we consider access, discovery, and search to be solved (although that is debatable) what is missing to enable humanities scholars to benefit more from the turn to LOD?

We invite scholars working on LOD projects within the larger spectrum of the humanities to participate in a workshop that aims to understand the limits of current work in this area.

All participants will be asked to draft a 1-page position paper (no citations or bibliography needed) that envisions a new LOD tool.

These papers will be shared before the event. Participants will rank the position papers that in their view are most likely to provide a means of advancing LOD for the humanities if taken up at the workshop. The four ranked highest will be asked to give one of the pecha-kucha-style pitches elaborating their position to kick off the day.

After these brief introductory talks and brief responses, participants will divide into groups, jumping off from these ideas to envision and share ideas for new tools, initiatives, and methods of working with linked data.

The working session groups will aim at developing collaboratively a fuller elaboration of the proposals to be released to the wider community, and working if feasible towards a plan for the collaborative development of one or more of the ideas.

At the end, all participants will reconvene to share ideas and summarize the results of each group. The notes for each session will be tidied up into blog posts and will then be collated into a white paper or report, along with the preliminary papers, that highlights the various gaps and opportunities within the LOD landscape.

Workshop leaders

Susan Brown

Susan Brown is a Canada Research Chair in Collaborative Digital Scholarship and Professor of English at the University of Guelph, and Visiting Professor at the University of Alberta. She researches Victorian literature, women's writing, and digital humanities. All of these interests inform *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*, an ongoing experiment in digital literary history published by Cambridge UP since 2006 that she co-directs. She directs the Canadian Writing Research Collaboratory, an online repository and research environment for literary studies in and about. Her current research touches on a range of topics in the digital humanities including interface design and usability, visualization and data mining, semantic technologies, and humanist-centered tool development. She is increasingly engaged with inquiry into how linked open data can serve humanities research. She also works on the impact of new technologies in the literature of the Victorian period. Brown is President of the Canadian Society for Digital Humanities/Société canadienne des humanités numériques.

Robert Warren

Rob Warren is a research associate with the Canadian Writing Research Collaboratory and Linked Modernisms projects at the University of Guelph. He also runs the Muninn Project, a large Linked Open Data project about the First World War. He is a professional engineer and adjunct professor in Mathematics and Statistics at Carleton University. Previously he was senior research fellow at the Big Data Institute in Halifax. His research interests lies in the design of ontologies and the application of artificial intelligence to digital humanities problems.

Abigel Lemak and Kim Martin

Abigel Lemak, a doctoral student at the University of Guelph, and Dr. Kim Martin, a Ridley Postdoctoral Fellow in Digital Humanities at the University of Guelph, will assist during the workshop.

Target Audience

Participants in DH2017 with a strong interest in linked data. We have invited a number of people who are very active in the field to join the workshop, and it looks like we will have input from the British Library, the Getty Institute, and the university library community, as well as DH researchers working from academic positions.

Let's Develop an Infrastructure for Historical Research Tools

Julia Luise Damerow
jdamerow@asu.edu
Arizona State University, United States of America

Dirk Wintergrün
dwinter@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science
Germany

Robert Casties
casties@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science
Germany

Scholars conducting historical research are provided with a growing range of digital humanities tools, supporting different phases of the research process: there is software for extracting text from documents (such as pdftotex, available on many Linux distributions or as part of Poppler), run OCR processes on images (for example Tesseract), tools for the creation, analysis, and visualization of datasets (for instance Nodegoat, Palladio, or Visualeyes), or software to work with annotations (for example Annotation Studio) or networks (such as Gephi or Cytoscape). Programming libraries are being developed to serve the needs of humanity scholars, like Spacy or Tethne. There are several repositories (such as HathiTrust or the Europeana) that provide access to sources and can easily be integrated into other services through APIs. Many tools, however, work well as self-contained units that scholars can use as singular parts of their research process, but cannot easily be combined into an integrated workflow by the researcher. Existing and new tools are developed using different languages and programming frameworks depending on requirements, skillset, and preference of the original developer, making reuse and integration harder for the developer seeking to combine several tools. Moreover, since most tools are developed independently of each other, many efforts are repeated by reimplementing functionality that is already provided by a different piece of software.

In this workshop, we would like to gather developers and programming-literate scholars to share their toolbuilding experiences and to present our first practical steps to create a system integrating multiple tools to work with historical documents from scan to analysis. The workshop is intended as a starting point for future exchange and cooperation for digital humanities developers.

In the summer of 2016, the Digital Innovation Group at Arizona State University (ASU) and the Max Planck Institute for the History of Science (MPIWG) started to combine their efforts in developing software for the history of science. One outcome of this collaboration is a research system that allows users to manage their documents, automatically runs OCR on uploaded files, provides an image viewer for uploaded and extracted images, and integrates document management with a multi-user Jupyter notebook server for writing analysis and visualization scripts. Rather than one big system, however, the research system is comprised of several integrated services developed independently of each other using different programming languages and frameworks.

For the Digital Humanities Conference 2017, we propose a full-day workshop with the goal to connect different tools and services to build a tool infrastructure for historical research.

The first half of the workshop will give tool developers a chance to present their software. Every presenter will be allowed 10 minutes for their presentation and 5 minutes for questions. ASU and MPIWG will present the different components of the developed research system. Specifically, we will present the following projects:

- Collaborative Jupyter Notebooks: a Jupyter Notebook server that allows sharing and publishing of notebooks based on Nextcloud and Dataverse.
- DocuManager: an environment for annotating, correcting and searching digital documents, in particular for OCR text in ALTO-XML and HOCR.
- Giles Ecosystem: an Apache Kafka-based service to extract images and texts from documents and run OCR procedures on them.
- Digilib: a Java-based IIIF-compliant image server and viewer.

The second half will be dedicated to discussing how different tools can be connected and integrated, and how we can build a community around those tools.

We envision the results of this workshop to be a concrete roadmap of how different tools will be integrated. We will define interfaces and API requirements, and if possible start development work during the workshop. Second, we will develop an organizational strategy for cooperation and collaboration among different projects. To aid organization, we will provide a Jira and Confluence project that participants can use during and after the workshop to organize collaboration.

We plan on organizing a follow-up meeting at the end of 2017 at Arizona State University to review progress since the initial workshop and plan next steps. If the collaboration is successful, we hope to establish regular meetings and expand the group to connect more tools and services.

Participants/Call

We will send out a call for participation in form of a short tool presentation for the first part of the workshop. We will ask presenters to focus on the technical perspective of their tool answering the following key questions:

- What is the general workflow of the tool?
- What is the core functionality of the tool?
- What input and output formats does the tool accept? Or what interfaces does it expose?
- What license model was chosen?
- What features are still missing and what are the next development goals?
- How is maintenance and development of the tool organized?

The deadline for the call is July 1st, 2017. We plan to accept 5-10 submission for our call, based on the usefulness of the tool and the potential for integration with other tools, which would all fit the first half of the workshop.

Audience

The target audience for this workshop are developers, historians with programming background, scholars with a technical background, and generally people involved in the development of tools to support historical research.

Confirmed Presenters

- Dirk Wintergrün (Max Planck Institute for the History of Science, Germany)
- Julia Damerow (Arizona State University, USA)
- Robert Casties (Max Planck Institute for the History of Science, Germany)
- Malte Vogl (Max Planck Institute for the History of Science, Germany)

High Performance Computing for Photogrammetry and OCR Made Easy

Quinn Dombrowski
quinnd@berkeley.edu
UC Berkeley, United States of America

Tassie Gniady
ctgniady@iu.edu
Indiana University, United States of America

Megan Meredith-Lobay
megan.lobay@ubc.ca
University of British Columbia, Canada

John Simpson
john.simpson@computecanada.ca
University of Alberta, Canada

Computationally-intensive research methods have seen increasing adoption among digital humanities scholars, but for scholars outside R1 institutions with robust computing environments, techniques like photogrammetry or text recognition within images can easily monopolize desktop computers for days at a time. Even at institutions with a research computing program, systems are configured for scientific applications, and IT staff may be unaccustomed to working with humanities scholars, particularly those who are not already proficient at using the command line. National compute infrastructures in North America (Compute Canada and XSEDE) are a compelling alternative, providing no-cost compute allocations for researchers and offering support from technical staff interested in and familiar with humanities computing needs. This workshop will start by introducing participants to Compute Canada and XSEDE, cover how to obtain a compute allocation (including for researchers outside of the US and Canada), and proceed through two hands-on tutorials on research methods that benefit from the additional compute power provided by these infrastructures: 1) photogrammetry using PhotoScan and 2) using OCR via Tesseract to extract metadata from images.

Photogrammetry

Photogrammetry (generating 3D models from a series of partially-overlapping 2D images) is quickly gaining favor as an efficient way to develop models of everything from small artifacts that fit in a light box to large archaeological sites, using drone photography. Stitching photographs together, generating point clouds, and generating the dense mesh that underlies a final model are all computationally-intensive processes that can take up to tens of hours for a small object to weeks for a landscape to be stitched on a high-powered desktop. Using a high-performance compute cluster can reduce the computation time to about ten hours for human-sized statues and twenty-four hours for small landscapes. Generating a dense cloud, in particular, sees a significant performance when run on GPU nodes, which are increasingly common in institutional HPC clusters and available through Compute Canada and XSEDE.

One disadvantage of doing photogrammetry on an HPC cluster is that it requires use of the command line and Photoscan's Python API. Since it is not reasonable to expect that all, or even most, scholars who would benefit from photogrammetry are proficient with Python, UC Berkeley has developed a Jupyter notebook that walks through the steps of the photogrammetry process, with opportunities for users to configure the settings along the way. Jupyter notebooks embed documentation along with code, and can serve both as a resource tool for researchers who are learning Python, and as a stand-alone utility for those who want to simply run the code, rather than write it. Indiana University, on the other hand, has developed a workflow using a remote desktop interface so that all the GUI capabilities and workflows of PhotoScan are still available. A python script is still needed so that the user may avail herself of the compute nodes, but the rest of the workflow is very similar traditional PhotoScan usage. Finally, both methods offload the processing the HPC cluster, allowing users to continue to work on a computer that might normally be tied up by the processing demands of photogrammetry.

The workshop will give participants hands-on experience creating a 3D model using two different approaches: first, by accessing the Photoscan graphical user interface on a virtual desktop running on XSEDE's Jetstream cloud resource; and second, by using a Jupyter notebook running on an HPC cluster.

OCR

Optical Character Recognition (OCR) is a tool used for extracting text from images and is perhaps most well known as a core technology behind the creation of the Google Books and HathiTrust corpora. OCR continues to open historical texts for analysis at large scale, fuelling a significant portion of research work within the digital humanities to the point that it would be difficult to think of the "million books problem" existing without this technology. While there are many OCR tools available the most popular tool that is also free and open source is Tesseract.

This portion of the workshop will also make use of Jupyter Notebooks to provide templates for learning the development process and that can then be taken away to speed development of future code. We will feature two projects for participants to practice with. A "traditional" OCR task that will have workshop participants processing images from the London Times in a demonstration of the improvements in OCR over the past few years and a task focusing on processing historical photographs to find text that can be added to the associated metadata to improve the searchability of an index.

Target Audience

We anticipate that this workshop will appeal particularly to scholars who work with cultural heritage materials (a field where photogrammetry is an increasingly common method for generating digital surrogates), as well as those who work with archival photographs, and scholars with large corpora of photographs. It will also be relevant for scholars who already engage in computational analysis of primary sources, who wish to increase the efficiency of their analysis by leveraging high-performance compute environments. No previous experience with HPC environments is necessary. This workshop can accommodate 25 participants.

Instructors

Quinn Dombrowski

Quinn is the Humanities Domain Expert at Berkeley Research Computing. At UC Berkeley, Quinn works with humanities researchers and research computing staff at Research IT to bridge the gap between humanities research questions and campus-provided resources for computation and research data management. She was previously a member of the program team for the Mellon-funded cyberinfrastructure initiative Project Bamboo, has led the DiRT tool directory and served as the technical editor of DHCommons. Quinn has an MLIS from the University of Illinois, and a BA and MA in Slavic linguistics from the University of Chicago.

Tassie Gniady

Tassie manages the Cyberinfrastructure for Digital Humanities group at Indiana University. She has a PhD in Early Modern English Literature from the University of California-Santa Barbara where she began her digital humanities journey in 2002 under the wing of Patricia Fumerton. She coded the first version of the NEH-funded English Broadside Ballad Archive, making many mistakes and learning much along the way. She now has an MIS from Indiana University, teaches a digital humanities course in the Department of Information and Library Science at IU, and holds regular workshops on text analysis with R and photogrammetry.

Megan Meredith-Lobay

Megan Meredith-Lobay is the digital humanities and social sciences analyst, as well as the Vice President, for Advanced Research Computing at the University of Briitsh Columbia. She holds a PhD from the University of Cambridge in medieval archaeology where she used a variety of computing resources to investigate ritual landscapes in early medieval Scotland Scotland. Megan has worked at the University of Alberta where she supported research computing for the Faculty of Arts, and at the University of Oxford where she was the programme coordinator for Digital Social Research, an Economic and Social Research Council project to promote advanced ICT in Social Science research.

John Simpson

John Simpson joined Compute Canada in January 2015 as a Digital Humanities Specialist and bringing a diverse background in Philosophy and Computing. Prior to Compute Canada, he was involved in a research-intensive postdoctoral fellowship focusing on developing semantic web expertise and prototyping tools capable of assisting academics in consuming and curating the new data made available by digital environments. He has a PhD in Philosophy from the University of Alberta, and an MA in Philosophy and BA in Philosophy & Economics from the University of Waterloo. In addition to his role at WestGrid, John is also a Member-at-Large of the Canadian Society for Digital Humanities (CSDH-SCHN), a Programming Instructor with the Digital Humanities Summer Institute (DHSI), and the national coordinator for Software

Séance d'initiation à la vidéo adaptée à la recherche SHS

Christian Dury
christian.dury@ish-lyon.cnrs.fr
Institut des Sciences de l'Homme, France

Cadrage de la problématique

L'image fixe ou animée est un outil de plus en plus important dans les recherches en sciences humaines et sociales. Son usage vient questionner le chercheur dans l'évolution du processus de recherche : recueil des données, analyse et restitution des résultats. Partant de ce constat, cette séance d'initiation se veut un espace de réflexion sur la place de l'image animée dans une recherche. Comment l'intégrer dans le travail scientifique ? Que peut-elle apporter au chercheur ? Quelles modifications relationnelles peut-elle entraîner dans le rapport au terrain ?

Dans une perspective de construction des savoirs, une vidéo permet de rendre compte de façon explicite des résultats de recherche en prenant plusieurs formes. Par exemple, elle peut être utilisée brute dans le cadre de recherche expérimentale, pour faire réagir à un phénomène par exemple. Elle peut se construire sous forme de documentaire pour expliciter, communiquer ou valoriser une théorie.

Qu'il s'agisse d'écrire un documentaire, de filmer une manifestation, d'enregistrer un entretien ou de capter une expérience, l'objectif de mise en image amène les chercheurs à s'interroger autour de la question : *« qu'est-ce que filmer en Sciences Humaines et Sociales ? »*.

Objectifs

Rien ne vaut la pratique pour se familiariser avec la technique audiovisuelle. Cette séance permet de découvrir les moyens d'analyser/restituer/communiquer/valoriser avec l'image animée autour de thématiques de recherche en SHS.

Cette séance d'initiation a pour objectif d'acquérir une autonomie avec le matériel sur des tournages légers et aborder l'ensemble des phases de réalisation (de l'écriture à la diffusion).

Lors de cette séance, les principales règles de captation vidéo seront abordées sous forme d'exercices pratiques autour de l'entretien filmée, par exemple. Ces images serviront de base pour aborder le montage numérique de données vidéos sur un logiciel de traitement des images et du son.

Prérequis

- Avoir des projets où l'image animée est mobilisée dans ses problématiques scientifiques.
- Avoir des envies d'utilisation de la vidéo dans ses recherches.

L'ISH organise ce genre de module d'initiation depuis juin 2016.

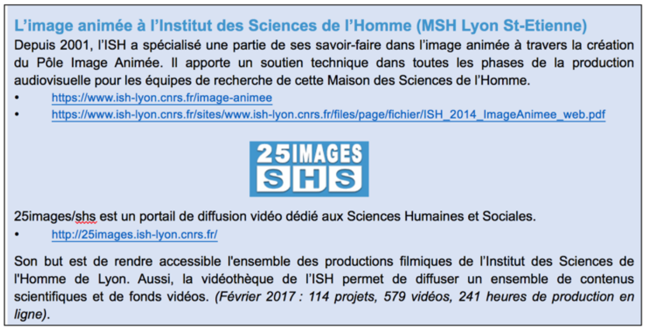
Aussi, en lien avec l'Ecole Doctorale « Sciences Sociales » de l'Université de Lyon 2, l'ISH co-organise les ateliers « Image dans les recherches en sciences humaines et sociales » (22h en 2017).

Programme de la journée

La matinée sera consacrée aux techniques de prise de vue adaptées à la recherche avec des exercices pratiques permettant d'approcher différents dispositifs filmiques.

L'après-midi s'articulera autour de la post-production en travaillant sur les prises de vue du matin sur un logiciel de montage.

La journée abordera enfin les questions de diffusion notamment avec l'expérience de la plateforme de diffusion vidéo de l'Institut des Sciences de l'Homme : 25images/shs .



L'image animée à l'Institut des Sciences de l'Homme (MSH Lyon St-Etienne)

Depuis 2001, l'ISH a spécialisé une partie de ses savoir-faire dans l'image animée à travers la création du Pôle Image Animée. Il apporte un soutien technique dans toutes les phases de la production audiovisuelle pour les équipes de recherche de cette Maison des Sciences de l'Homme.

- https://www.ish-lyon.cnrs.fr/image-animee
- https://www.ish-lyon.cnrs.fr/sites/www.ish-lyon.cnrs.fr/files/page/fichier/ISH_2014_ImageAnimee_web.pdf

25images/shs est un portail de diffusion vidéo dédié aux Sciences Humaines et Sociales.

- http://25images.ish-lyon.cnrs.fr/

Son but est de rendre accessible l'ensemble des productions filmiques de l'Institut des Sciences de l'Homme de Lyon. Aussi, la vidéothèque de l'ISH permet de diffuser un ensemble de contenus scientifiques et de fonds vidéos. *(Février 2017 : 114 projets, 579 vidéos, 241 heures de production en ligne).*

The Design of Historical Data Projects: The Comédie Française Registers Project and the Laboratoire Paris XVIII. A hands-on workshop and conversation about creating datasets for historians and developers.

Jamie Folsom
jamie@performantsoftware.com
Performant Software Solutions LLC
United States of America

Pascal Bastien
bastien.pascal@uqam.ca
Université de Quebec à Montréal, Quebec, Canada

Jeffrey Ravel
ravel@mit.edu
Massachusetts Institute of Technology
United States of America

Sara Harvey
saraharvey@uvic.ca
University of Victoria, United States of America

Julian Puget
puget.julien@gmail.com
Université de Quebec à Montréal, Quebec, Canada

Benjamin Deruelle
deruelle.benjamin@uqam.ca
Université de Quebec à Montréal, Quebec, Canada

Abstract

The Comédie Française Registers Project (or “CFRP”) and the Laboratoire Paris XVIII *(*or “LP18”) both aim at understanding social and cultural phenomena in Paris in the 18th century by collecting and utilizing data from relevant historical documents, and creating and supporting communities of practice and scholarship around those data using web technologies.

Through an overview of the Comédie Française Registers Project (CFRP), and hands-on engagement with the dataset compiled by that project, participants will gain a shared context for discussion of historical data projects.

On that basis, we will introduce the “Laboratoire Paris XVIII” (LP18), which aims to create a collaborative workspace to support the compilation of datasets about life in 18th-century Paris from primary source materials, and visualizations of those data in time and space.

Participants in the workshop will then engage in a discussion of how the “LP18” project and platform should be designed, in light of the successes of the CFRP project, to maximize its accessibility and utility.

This workshop is aimed at researchers and developers, and presented in French and English.

Description

Part 1: Introduction: The Comédie Française Registers Project

We’ll open with a presentation of the Comédie Française Registers Project by two of the project’s principal investigators. From the CFRP website:

> *From 1680 until 1791, only one theater troupe in Paris was allowed to perform the plays of Molière, Corneille, Racine, Voltaire, Beaumarchais, and every other French-language playwright. This troupe, the Comédie-Française, played the works of these authors over 34,000 times in this period. Remarkably, the troupe kept detailed records of their box office receipts for every single one of those performances. These daily receipt registers, still housed today in the troupe’s archives in the heart of Paris, are now available online via the Comédie-Française Registers Project.*

The CFRP has created web-based tools to collect and manage ticket receipt data, and to search, analyze and visualize those data to support research and collaboration.

We will demonstrate some examples of how scholars and developers have integrated the CFRP data into research and teaching practice, extended it for specific use cases, connected with other datasets, and contributed their work back to a growing community of practitioners.

- Project overview, history, and results to date: Dr. Jeff Ravel, Professor and Head of History at MIT
- Tools, visualizations, and future work planned: Dr. Sara Harvey, Professor of French Literature at University of Victoria

Part 2: Hands on with CFRP data

We will then move to structured hands-on work with the CFRP dataset, in which participants may try out the various modalities of access to the data, come to grips with the data themselves, ask and answer research questions, and envision new applications.

After an orientation to the CFRP website, tools and applications, participants will select from a several structured,practical activities to be done individually or in pairs. Members of both the CFRP and the LP18 project teams will participate and answer questions.

Part 3: Le Laboratoire Paris XVIII

Next, we will introduce the LP18 project, with an overview of the project's goals, context and methods. From the project description:
Labo Paris XVIII, or LP18, and the software it envisions, will be a platform for assembling, sharing and using datasets relevant to life in Paris in the 18th century. These datasets will provide lenses through which to view the "information networks" of the time, and through them, the life of the city and its citizens. By assembling a diversity of archival material, most of it unpublished, LP18 aims to make it possible to search a unique corpus of sources, some printed, others handwritten, and to visualize them on an historical map of the city. Members of the project team will comment on key aspects of the project.

- Historical context and project overview: Dr. Pascal Bastien
- Geospatial considerations: Dr. Julien Puget
- Environmental Scan: Dr. Benjamin Deruelle

Part 4: Discussion

Following the introduction of the LP18 project, will be a discussion of the fields of research and technology in which these two projects are situated, with the aim of informing how LP18 can be informed by those other efforts, and designed for accessibility and utility.
Presenters will frame key questions to introduce the conversation. Topics may include:

- Practical considerations in collaborative data projects
- Document digitization, ingestion and data extraction
- Intellectual property considerations in historical data projects
- Example projects, datasets, tools and software
- Database and API design
- Using open source software and open standards, production open data
- Communicating across disciplines, languages and time zones

Conclusion

Time will be reserved at the end of the workshop for feedback, thoughts and questions.

Digital Scholarship and Privacy-sensitive Collections

Unmil Karadkar
unmil@ischool.utexas.edu
University of Texas at Austin, United States of America

King Davis
king.davis@ischool.utexas.edu
University of Texas at Austin, United States of America

Introduction

Humanities scholars have historically used archives that include restricted or privacy-sensitive collections in order to conduct their investigations about sensitive topics. The recent developments in digitization and dissemination technologies present the possibility of making archival collections broadly available. Furthermore, collections of new, born-digital documents will be readily available to support and enhance scholarship. However, such access has also exacerbated threats to the privacy of individuals named in these records. Examples of such privacy-sensitive records include mental health institutional records, prison records, records of the Truth and Reconciliation Commission, Nazi archives, and the Guatemalan police archives. In the physical world, access to these records is protected by distance, physical access, and a variety of national and local statutes. The legal framework for digital records is substantially behind that for physical records. Furthermore, the online availability of such records has a potential to stigmatize or embarrass the families or descendants of those named in the records when they bear no responsibility for the acts or afflictions of the named individuals, raising ethical issues in providing broad, open access to these records.

The organizers are studying the legal, conceptual, and practical issues in harnessing such privacy-sensitive collections in the service of scholarship—for example in history (of medicine and mental health), law, and social services. Our research is funded by the Andrew W. Mellon Foundation (grant number: 11500653) under the scholarly communications program.

This workshop will invite broad participation from scholars and practitioners who work with or are interested in issues surrounding humanities scholarship supported or enhanced by digital, privacy-sensitive collections. A non-exhaustive list of topics of interest include:

- Digitization, curation, and preservation of privacy-sensitive collections
- Theoretical and metadata models
- Policies, workflows, and protections for accessing materials

- Issues in using cloud services for privacy-sensitive materials storage and scholarship
- Scholarly information behavior and needs
- Models for balancing privacy of named entities versus access to specific demographics
- Mechanisms and models for data retrieval from handwritten documents
- Privacy-aware digital repository architectures
- Privacy-aware crowdsourcing and transcription methods
- Privacy issues in designing user interfaces and data visualizations
- Privacy mitigation in data analytics and presentation
- Evaluation of existing software, infrastructure, and techniques
- Social justice issues and non-scholarly outcomes of work with restricted collections

Workshop organizers

Unmil P. Karadkar

Unmil P. Karadkar is an assistant professor in the School of Information at The University of Texas at Austin. He situates his work at the intersection of digital libraries, human-computer interaction, and visualization. He studies data practices of researchers with an eye toward identifying unmet information needs. Based on an understanding of these needs, he designs software to support their evolving practices and evaluates the impact of this software on their work. His research contributes to areas such as the design of digital collection interfaces and digital humanities. His research has been funded by the National Science Foundation, Texas General Land Office, USAA, and most recently, the Andrew W. Mellon Foundation.

King Davis

King Davis is a research professor in the School of Information at The University of Texas at Austin and has made outstanding contributions in the field of Health and Mental Health over the last three decades. Dr. Davis held the Robert Lee Sutherland Chair in Mental Health and Social Policy at the University of Texas at Austin, School of Social Work. From 2003-2008, Dr. King also served as the Executive Director of the Hogg Foundation, which awards grants and manages programs to improve mental health research and services in Texas. Prior to his work in Texas, Dr. Davis served as the Commissioner of the Department of Mental Health, Mental Retardation and Substance Abuse Services for the Commonwealth of Virginia by Virginia Governor L. Douglas Wilder. He also has served as the John Galt Chair in Public Mental Health at the University of Virginia's Department of Psychiatry. Dr. Davis has held academic appointments at Washington University in St. Louis, Virginia Commonwealth University, Eastern Virginia Medical School and Norfolk State University. Dr.

Davis received his PhD from Brandeis University, Florence Heller School for Social Policy and Management, and his MSW from California State University Fresno, School of Social Work. He has written and published numerous articles and reports on mental health, fund raising, managed health care and social justice. His book, The Color of Social Policy, was published in 2004.

Target audience and expected number of participants

This is our first workshop on this topic and we are unsure what the interest will be. In this light, we have tried to make the topic as broad as possible, while retaining the core characteristics and constraints that we have found in our work. Our poster at DH 2014 was well-received and several conference attendees expressed an interest as well as challenges in working with collections similar to ours. The collections they described were geographically and topically diverse. With this experience, we anticipate receiving 10 to 20 submissions and accepting 8 to 10 for presentation at the workshop. We are unable to gauge the level of interest in the DH community in attending the workshop without presenting.

Length and format

We propose to hold our workshop for one day. This time frame will allow for adequate exploration of the various aspects of the workshop topics as well as domains via presentations, as well as in-depth discussion. We are open to conducting the workshop on a single day or as two half days.

The workshop will be held in a seminar style, with several short and long presentations. Individuals may participate in the workshop without presenting. The organizers will include an open discussion time to engage the audience and, especially, to tease out aspects of scholarship with privacy-sensitive digital collections that the presentations do not cover.

Program Committee

We will assemble a diverse program committee that includes scholars and practitioners with a diverse expertise. Potential invitees include:

- Tom Cramer, Assistant University Librarian-Library Technology, Stanford University
- Karen Estlund, Associate Dean for Technology and Digital Strategies, Penn State University Libraries
- Donald Fyson, Professor-History, Universite Laval
- Pat Galloway, Professor-Information Studies, The University of Texas at Austin Gary Geisler, UX Designer, Stanford University Libraries
- Geoffrey Rockwell, Professor-Philoshophy and Humanities Computing, University of Alberta Martin Summers, Associate Professor-History, Boston College

We welcome input from the DH 2017 program committee for additional suggestions for program committee members.

CATMA 5.0 Tutorial

Jan Christoph Meister
jan-c-meister@uni-hamburg.de
University of Hamburg, Germany

Evelyn Gius
evelyn.gius@uni-hamburg.de
University of Hamburg, Germany

Jan Horstmann
jan.horstmann@uni-hamburg.de
University of Hamburg, Germany

Janina Jacke
janina.jacke@uni-hamburg.de
University of Hamburg, Germany

Marco Petris
marco.petris@uni-hamburg.de
University of Hamburg, Germany

What is CATMA?

This hands-on tutorial introduces humanists to CATMA (Computer Aided Text Markup and Analysis), a tool developed at the University of Hamburg and currently used by over 60 research projects worldwide. CATMA offers a unique combination of three main features found in no other text analysis tool:

CATMA supports **collaborative annotation and analysis** – a text or text corpus can be investigated individually, but also jointly by a group of students or researchers.

CATMA supports **explorative, non-deterministic practices of text annotation** – a discursive, debate-oriented approach to text annotation based on the research practices of hermeneutic disciplines is the underlying conceptual model.

CATMA integrates **text annotation and text analysis** in a web-based working environment – which makes it possible to combine the identification of textual phenomena with their investigation in a seamless, iterative fashion.

What sets CATMA apart from other digital annotation methods is its 'undogmatic' approach: the system does neither prescribe defined annotation schemata or rules, nor does it force the user to apply rigid yes/no, right/wrong taxonomies to texts (even though it allows for more prescriptive schemata as well). Rather, CATMA's logic invites users to explore the richness and multi-facettedness of textual phenomena according to their needs: Users can create, expand, and continuously modify their own individual tagsets – so if a text passage invites more than one interpretation, nothing in the system prevents assigning multiple, or even contradictory annotations. Despite all this flexibility, CATMA does not produce idiosyncratic annotations: All markup data can be exported in TEI/XML-format and reused in other contexts.

Since CATMA is a highly intuitive tool, it is also suitable for humanists with little technical knowledge: the GUI allows for a quick kick-off, and CATMA's query builder (a step-by-step dialogue-based widget) helps users retrieve complex information from texts without having to learn a query language. Another plus on the easy-to-use side is the fact that CATMA's automated distant-reading functions are continuously enhanced and extended – the current version 5.0 already features a number of automated annotation routines, among others the identification of basic narrative features in texts.

The aim of the tutorial

In our half-day tutorial, we will introduce the core annotation and analysis functionalities of CATMA and show how they can be combined with the annotations provided automatically. Participants will be taken in a step-by-step, hands-on approach through the full cycle of a CATMA-based text investigation:

- From text upload to initial textinvestigations,
- then to annotation and specification of annotation categories,
- from there to combined text queries that consult the source text and its annotations in combination,

and finally to the visual output of queryresults.

In a later phase of the tutorial, participants will have the opportunity of testing the tool with regards to their own research interests: They can annotate their own texts or annotate collaboratively a text we will provide. We would also like to engage participants in a critique of CATMA's design and components as well as a general discussion about requirements for text analysis tools in their fields of interest.

The primary users of CATMA are literary scholars, as well as graduate and undergraduate students of Literary Studies. Nevertheless, this tutorial is likely to be of interest also to:

humanities scholars of ALL fields concerned with text analysis (with and without experience in digital text analysis),

software developers in the humanities interested in non-deterministic text analysis and automated annotation.

Participants need no prior knowledge of digital text annotation and can work with their own laptop computers and their own digital texts. CATMA runs on Laptop or PC (Windows, Unix or MacOS) with a current web browser (MS Explorer or Edge; Firefox, Chrome, Safari) with a mouse or touchpad. Touchscreen navigation is not yet supported (but in the pipeline!). The room in which the workshop takes place should accommodate 25–30 people and provide WLAN and a projector.

Tutorial Instructors

All tutorial instructors come from the developing team of the CATMA project and/or the forTEXT project that is building a platform starting from CATMA. We have been presenting and teaching CATMA on various national and international occasions in the last years.

Evelyn Gius

Evelyn is working in the field of Digital Humanities as a researcher and has been involved in the creation of CATMA from the very beginning. Her research focus is on manual and automated text analysis. For her PhD project in Literary Studies she has explored with CATMA the benefits of applying narratological categories from literary studies to the analysis of narrations of labor conflicts.

Jan Horstmann

Jan uses CATMA as a tool for textual analysis in literature studies and narratology. Currently he is investigating works of Goethe with a combination of distant and digital close reading methods. His focus is to improve the usability of digital tools for people with little or no prior knowledge of computing or programming, i.e. researchers from classical literature studies.

Janina Jacke

Janina Jacke has worked in the heureCLÉA project (2013–2016) that was aimed at developing automated annotation routines for CATMA. Since her research focuses on narratology and theory of interpretation, her main interest in the DH-context lies in working out the theoretical prerequisites for automated literary annotation.

Jan Christoph Meister

Chris is Professor of German literature with a main research focus in the Digital Humanities. As original inventor of CATMA, he has led several projects concerned with the annotation and visualization of literary data and the development and enhancement of DH-tools.

Marco Petris

Marco is a computer scientist with a strong affinity for the humanities and has been engaged in the creation of CATMA from the very beginning. As a research developer he is involved in all aspects of the design and implementation of tools for the Digital Humanities.

Hands on Text Analytics with Orange

Ajda Pretnar
ajda.pretnar@fri.uni-lj.si
University of Ljubljana, Slovenia

Niko Colnerič
niko.colneric@fri.uni-lj.si
University of Ljubljana, Slovenia

Lan Žagar
lan.zagar@fri.uni-lj.si
University of Ljubljana, Slovenia

Orange for Text Analytics

In recent years, the digital humanities community has been introduced to many powerful tools for text analysis, but few of these tools combine powerful data mining and machine learning algorithms within a simple and capable user interface. For flexible and creative analysis, researchers need a tool that focuses on intuition, visualizations and interactivity.

This workshop will introduce participants to Orange, a visual programming environment for data mining, suitable for both beginners and experts. Particular emphasis will be placed on its Text add-on, which offers components for text mining, visualization and deep- learning-based embedding.

This is a hands-on workshop, where the participants will actively construct analytical workflows and go through case studies with the help of the instructors. They will learn how to manage textual data, preprocess it, use machine learning, data projection and visualisation techniques to expose hidden patterns and evaluate the resulting models. At the end of the workshop, the participants will know how to use visual programming to seamlessly construct powerful data analysis workflows, which can be applied to a wide range of challenges in digital humanities.

Structure of the Workshop

Part 1: Visual programming, workflows, data input and preprocessing

First, we will show the basics of Orange: how to load the data, inspect and visualize it. Participants will be introduced to several options for data import, from standard Corpus to Twitter, Guardian and Text Import. Once the corpus is loaded, we will preprocess it and display the result in a word cloud. A particular emphasis will be on the use of custom preprocessing techniques and how to successfully apply them to the corpus. The results of each technique will be observed in an interactive word cloud and concordances.

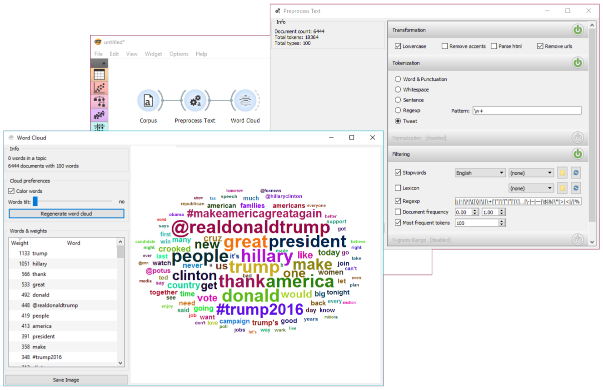


Figure 1: Preprocessing results displayed in a word cloud

Part 2: Machine learning and deep-learning-based embedding for predictive analysis

Next, we will use Twitter data to construct an author prediction pipeline and test some classifiers. We will fetch author Timelines from Twitter and observe the retrieved corpus. This time we will introduce a pre-trained tweet tokenizer and pass the preprocessed corpus through a bag of words. We will discuss bag of words parameters and how to best prepare the data for further analysis. The results of using different parameters will be observed in a data table to understand the underlying data structures. For comparison, we will use deep-learning-based embedding to derive vector representation of tweets and in this way enable machine learning.

We will explain how we can use machine learning in text mining and introduce a number of techniques for predictive analysis. We will use cross-validation to test the constructed bag of words models and compare classification scores for each algorithm. We will discuss the quality of constructed models and what scores are usually the best for observing model quality. Additionally, we will inspect misclassified tweets in a confusion matrix and even further in Corpus Viewer, to leverage the possibilities of a close(r) reading.

Part 3: Data clustering, sentiment analysis, image and geo analytics

In the third part, we will work on geomapping and image analytics. We will transform textual and visual data into feature vectors and plot these data onto a world map to discover interesting relations.

We will discuss how to acquire geolocated data from Twitter and why this is useful. Next, we will use geotagged Twitter data and apply a pre-trained sentiment analysis model to acquire sentiment orientation. We will map the sentiment-tagged tweets and explore how to use sentiment together with geomapping.

Finally, the participants will be introduced to image analytics for humanities research. We will explain why and how to transform raw images into multidimensional vectors and how to work with the new data. We will cluster Instagram images into groups and explore how to map image-containing tweets on a world map. Do images correspond to geolocation? We will see.

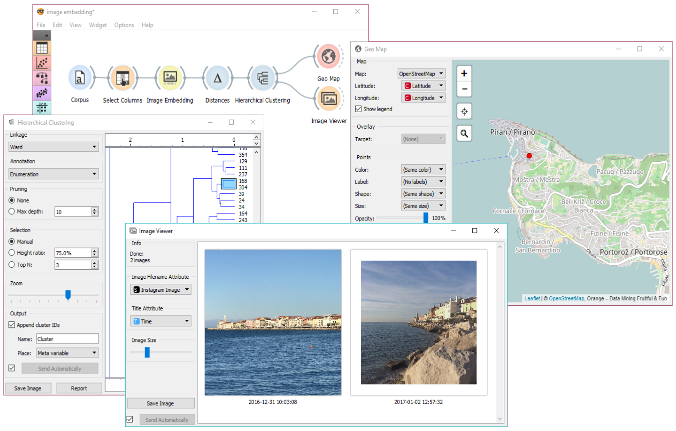


Figure 2: Images from social media are embedded with ImageNet embedding, clustered with Hierarchical Clustering and displayed on a map by their geolocation.

From Text to Networks: Combining Entity and Segment Annotations in the Analysis of Large Text Corpora

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
University of Stuttgart, Germany

Maximilian Overbeck
maximilian.overbeck@sowi.uni-stuttgart.de
University of Stuttgart, Germany

Sandra Murr
sandra.murr@ilw.uni-stuttgart.de
University of Stuttgart, Germany

Introduction

In this half-day tutorial we will offer a full-fledged, implemented and tested workflow that has been developed in the interdisciplinary *Center for Reflected Text Analytics* (CRETA, a research center connecting both scholars from Humanities/Social Sciences and Computational Linguistics at the University of Stuttgart). Our focus is the valid and reliable identification of various kinds of entities and segments from raw, un-annotated texts and the extraction of specific relational information via network visualizations. Given the recent interest in networks for data representation and visualization (e.g., Gephi-tutorial at DH 2016), we argue that the following three-step-workflow is applicable to many research questions in the Social Sciences and Humanities:

1. Detection of entity references in texts of different genres (*e.g. references to chancellor Merkel in parliamentary debates)*,
2. Segmentation of the texts guided by research questions (*e.g. parts of a parliamentary speech dealing with the Greek financial crisis*), and
3. Creation of networks of entities that co-occur within a segment *(e.g. references to national or international organizations in a parliamentary debate dealing with the issue of wars and military interventions).*

This workflow is one example of modularizing complex research questions into concrete steps and can moreover be combined with computational methods for the semi-automatic analysis of very large text corpora. The concepts of "entity" and "segment" are sufficiently generic to allow the same set of tools to be employed in different research questions originating from different fields of research. The tutorial is therefore not aimed at a specific Humanities or Social Sciences discipline and instead open to all researchers interested in the analysis of entity relations in large amounts of textual data.

In our tutorial we will make use of the web-based annotation tool CRETAnno developed to support semi-automatic annotation. CRETAnno provides tools for annotation and continuous assessment of *inter-annotator agreement*, thereby facilitating the production of reliable and valid data. Our tool facilitates the annotation of large text corpora: After some training instances are annotated, a machine learning model can be trained to predict new instances on additional texts, which can then be corrected and used as additional training material. This way, large texts can be annotated (relatively) quickly, given systematic manual annotation and clear annotation guidelines. This 3-step approach is currently investigated within the *Center for Reflected Text Analytics* (CRETA) on four distinct text corpora, connected to diverse research questions in different disciplines. Although establishing broadly applicable workflows has its merits (Kuhn & Reiter, 2015), we believe it is important to be able to "parameterise" them to take into account the specificities of a concrete research question. Research questions should govern the definition of entities, segments and weighting criteria in the network. In the tutorial, participants will be free to bring in (and work on) their own research questions (within the time limits of the tutorial).

Entity Reference Detection

Every concept of interest within a real or fictional world can be considered as an entity. Words in a text refer to th ese entities and are therefore called *entity references*. We have established annotation guidelines that distinguish six entity classes, oriented at the research questions within CRETA: *Person*, *Location*, *Organization*, *Work* (e.g., a piece of art), *Event* and *Abstract Concept* (e.g., art).

While these entities are semantically diverse, their linguistic representation in texts is similar: References are either proper nouns (*Hillary Clinton*/*EU*), pronouns (*she*/*it*) or appellative noun phrases (*an American politician* / *the international organisation*).Most of the entity references consist of a few words, but we generally opt for annotating full noun phrases (e.g., *the British people after having voted for the Brexit*). In order to be able to link entities semi-automatically, we focus on appellative noun phrases and proper nouns, and ignore pronouns (see below).

The notion of "entity reference" we are aiming for differs from what is known in **Named Entity Recognition** (NER) and **Coreference Resolution** (CR). In NER, only proper nouns are detected, while CR also aims to resolve pronouns. Our notion of entity reference detection is aiming for the middle ground. By excluding pronouns, we also exclude the most ambiguous words, whose co-reference properties typically can only be judged in context of their appearance. Appellative NPs contain enough information such that we can establish their identity with proper nouns with relatively simple lists and rules.

Text Segmentation

Researchers from Humanities or Social Sciences generally want to inquire either the interaction between entities (within certain contexts) or between entities and the contexts themselves. Text segmentation is our way of operationalising this context. The notion of segment -- again -- is a generic one, to be adapted to specific research questions and/or theoretical assumptions made within a discipline or research area. Different kinds of segmentation are distinguishable: A **segmentation according to structural units** like chapters (narratives), speeches (minutes of parliamentary debates) or acts (dramatic texts) relies on the proper detection of such segments in the original texts and is therefore highly intertwined with the concrete text format at hand. Although machine learning models can be trained to perform such tasks, they likely do not generalize well to new texts. Even in TEI-encoded dramatic texts (which are strongly structured), there are a lot of options how to encode acts. We therefore aim for making it easy for researchers from Humanities and Social Sciences to detect such segments using metadata (e.g. dates of publication of a newspaper article or a parliamentary debate), text-specific regular expressions and/or rules.

A second kind of segmentation is **segmentation according to content** criteria. Depending on text genre and research question, this can mean segmentation by topic, narrative level, plot, time, location etc. One possible application is the segmentation of newspaper content according to various topics (Kantner & Overbeck 2017, forthcoming).

Structurally, segment annotations differ from entity reference annotations by being longer and thus sparser within a text. This has consequences for the semi-automatic support, because annotating a sufficient number of training instances requires more text to be read (and analysed with respect to its segmentation) and thus takes more time. CRETAnno therefore supports a number of unsupervised segmentation algorithms that can be used directly. In addition, researchers can specify text patterns using regular expressions and simple rules and thus focus the segmentation on the specific research question they have.

Entities + Segments = Networks

Given entity reference and segment annotations, it is only a small step to extract network-like data based on co-occurrence. As the entity reference annotation does not include links between annotations referring to the same entity, we developed a small tool to mark co-reference, given the annotated entity references. Currently, this has to be done manually, but we will explore automatisation possibilities in the future. Given that we can already identify string-identical entity references automatically, it is a manageable workload.

CRETAnno offers an interface to the graph exploration software Gephi, which can be used to edit, explore, inspect and visualise the network (the tutorial covers the annotation, ex- and import, but only basic functionality of Gephi.).

Tutorial

Participants will have the opportunity to work on texts of their own choosing within the first half of the workshop. To that end, they will be asked to submit their texts before the workshop. We will supply hands-on material to participants that do not submit. The tutorial focuses on hands-on sessions and active participation.

Appendix

Tutorial Instructors

All submission authors work jointly in the Center for Reflected Text Analytics (CRETA) at Stuttgart University, Germany.

Sandra Murr

Sandra Murr , is a PhD candidate in the Department of modern German literature at the University of Stuttgart. Within CRETA, she analyzes literary works of the productive reception of J. W. v. Goethe's *Sorrows of the Young Werther*, the so-called Wertheriaden, focusing on the analysis of the central character constellation with respect to emotions.

Maximilian Overbeck

Maximilian Overbeck is a PhD candidate in Political Science at the Chair of International Relations and European Integration at the University of Stuttgart. In his PhD he analyses Western debates on religion in the context of wars and armed conflicts where he uses highly innovative computational-linguistic approaches for the valid and reliable analysis of large newspaper corpora.

Nils Reiter

Dr. Nils Reiter works at the Department of Natural Language Processing and coordinates the scientific work in CRETA. Since his PhD thesis with the title Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms (Link), he is working in and for the Digital Humanities area, with a particular focus on literary texts, annotation and the operationalisation of Humanities research questions.

Target Audience

Any student or scholar interested in qualitative and quantitative text analysis is invited. Prior knowledge in text analysis techniques is not obligatory but might be helpful. Programming skills are not necessary, but familiarity with Gephi is helpful. We welcome 20 to 30 participants.

Acknowledgements

We at CRETA are grateful to the German Federal Ministry of Education and Research (BMBF) for its generous funding in the years 2016 until 2018 (project ID: 01UG1601).

Bibliography

John, M., Lohmann, S., Koch, S., Wörner, M., Ertl, T. (2016) Visual Analytics for Narrative Text: Visualizing Characters and their Relationships as Extracted from Novels. *Proceedings of the 7th International Conference on Information Visualization Theory and Applications* (IVAPP '16). SciTePress, 2016.

Kuhn, J., and Reiter, N. (2015). A Plea for a Method-Driven Agenda in the Digital Humanities. In *Proceedings of Digital Humanities 2015*, Sydney, Australia, June 2015.

Kantner, C., Overbeck, M. (2017, forthcoming): „Die Analyse ‚weicher' Konzepte mit ‚harten' korpuslinguistischen Methoden. In: J. Behnke, A. Blaette, J.-U. Schnapp & C. Wagemann (eds.) *Big data? New Data*. Baden-Baden: Nomos Verlag.

Overbeck, M., (2015). Observers turning into participants: Shifting perspectives on religion and armed conflict in Western news coverage. The Tocqueville Review/La revue Tocqueville, 36, 95-124.

Reiter, N. (2015) Towards Annotating Narrative Segments. Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 34–38, Beijing, China, July 30, 2015.

Transkribus: Handwritten Text Recognition technology for historical documents

Louise Seaward
louise.seaward@ucl.ac.uk
University College London, United Kingdom

Maria Kallio
maria.kallio@arkisto.fi
National Archives of Finland, Finland

Topic

Transkribus is a platform for the automated recognition, transcription and searching of handwritten historical documents. Transkribus is part of the EU-funded Recognition and Enrichment of Archival Documents (READ) project. The core mission of the READ project is to make archival material more accessible through the development and dissemination of Handwritten Text Recognition (HTR) and other cutting-edge technologies.

The workshop is aimed at researchers and students who are interested in the transcription, searching and publishing of historical documents. It will introduce participants to the technology behind the READ project and demonstrate the Transkribus transcription platform. Our team has already conducted 30 similar workshops over the course of 2016, including several sessions with digital humanities scholars and students.

Transkribus can be freely downloaded from the Transkribus website. Participants will be instructed to create a Transkribus account and install Transkribus on their laptops in advance of the workshop. They should bring their laptops along to the workshop.

The workshop will consist of five parts:

Introduction to Handwritten Text Recognition (HTR) technology

The introduction to this workshop will explain how new algorithms and technologies are making it possible for computer software to process handwritten text. Handwritten Text Recognition (HTR) technology works differently from Optical Character Recognition (OCR) for printed texts (Leifert et al., 2016). Rather than focusing on individual characters, HTR engines process the entire image of a word or line, scanning it in various directions and then putting this data into a sequence. This introduction will outline the workings of HTR technology and show examples of the successful automatic transcription and searching of historical documents. It will also explain the possibilities of working with different languages and styles of handwriting. The latest experiments demonstrate that Transkribus can automatically generate transcripts with a Character Error Rate of 5-10%. This means that 90-95% of the characters in the transcript would be correct.

Overview of the READ project

This presentation will give an overview of the READ project and the specific tools it is creating. Computer scientists working on READ are developing HTR technology using thousands of manuscript pages with varying dates, styles, languages and layouts. Testing the technology on a large and diverse data set will make it possible for computers to automatically transcribe and search any kind of handwritten document, from the Middle Ages to the present day, from old Greek to modern English. This research has huge implications for the accessibility of the written records of human history. The READ project is making this technology available through the Transkribus platform but is also developing other tools designed to make it easier for archivists, researchers and the public to work with historical documents. The workshop leaders will present prototypes of some of these tools. These include a system of automatic writer identification, an e-learning app to enable users to train themselves to read a particular style of writing, a mobile app to allow users to digitise and process documents in the archives and a crowdsourcing platform where volunteers can transcribe with the assistance of HTR technology. These tools will be open source and are designed to be used and adapted by other institutions and projects.

Introduction to Transkribus

HTR technology is made available through the Transkribus platform, which is programmed with JAVA and SWT (Mühlberger et al.) A transcription of a handwritten document can be undertaken in Transkribus for two main purposes. The first is a simple transcription – this allows users to train the HTR engine to automatically read historical papers. The second is an advanced transcription – this allows users to create a transcription of a document which may serve as the basis of a digital edition. This presentation will explain both uses of Transkribus.

HTR engines are based on algorithms of machine learning. The technology needs to be trained by being shown examples of at least 30 pages of transcribed material. This helps it to understand the patterns which make up words and characters. This training material is known as 'ground truth' (Zagoris et al., 2012, Gatos et al., 2014). The workshop leaders will demonstrate how 'ground truth' training data can be prepared using Transkribus. Participants can work with images of their own documents, or experiment with test documents already on the system.

Transkribus can also be used simply for transcription. This presentation will explain how to create a rich transcription of a document in the platform, using structural mark-up, tagging, document metadata and an editorial declaration.

Working independently with Transkribus

In the last part of the workshop, the participants will be able to try out the functions of Transkribus on their own laptops. They will be supported by the workshop leaders, who will explain the different elements of the platform and then give participants the chance to practice each function for themselves. The workshop leaders will circulate around the room to answer any questions.

The workshop leaders will demonstrate the following tasks. After each demonstration, participants will be given 10-15 minutes to practice what they have learned.

- Document management – how to upload, view, save, move and export documents in standard formats (PDF, TEI, docx, PAGE XML)
- User management – how to allow specific users to view and edit documents
- Layout analysis – how to segment your documents to create training data for the HTR engines
- Transcription – how to create a rich transcript with tags and mark-up
- HTR – how to apply HTR models to automatically generate transcripts, how to conduct a keyword search of your documents, how to assess the accuracy of automatically generated transcripts

Question and Answer

The workshop will close with a Question and Answer session where participants can clarify anything they are unsure about. They will also have the opportunity to provide feedback on the Transkribus tool via our user survey.

Organizers

Louise Seaward

Dr. Seaward received her PhD in History from the University of Leeds (United Kingdom) in 2013. She is currently a research associate at University College London where she coordinates 'Transcribe Bentham', the scholarly crowdsourcing initiative which asks members of the public to transcribe manuscripts written by the British philosopher Jeremy Bentham (1748-1832). Outside of digital humanities and Bentham, her research interests relate to the history of censorship and the Enlightenment.

Maria Kallio

Maria Kallio works as a Senior Research Officer at the National Archives of Finland where she is responsible for collections, crowd-sourcing and dissemination within the READ project. Currently she is also finishing her PhD in History at the University of Turku (Finland). In addition to digital humanities, her research interests include medieval literacy and written culture in all its diversity.

Proposed audience

Humanities and digital humanities scholars, archivists, librarians, computer scientists

Guidelines for Participants

Participants should register to attend the workshop by sending an email to Louise Seaward. Participants will need to bring their own laptops and install Transkribus before attending the workshop. If participants are interested in working with their own documents, they should bring a selection of digital images to the workshop. Otherwise, it will be possible to work with test documents already on the platform.

Bibliography

Leifert, G., Strauß, T., Grüning, T., and Labahn, R. (2016). 'Cells in Multidimensional Recurrent Neural Networks', https://arXiv.org/abs/1412.2620v02

Mühlberger, G., Colutto, S., Kahle, P., (forthcoming) 'Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP)' (pre-print)

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J.A., Toselli, A.H., and Vidal, E. (2014). 'Ground-Truth Production in the tranScriptorium Project', Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on Document Analysis Systems, 237-244

Stamatopoulos, N., and Gatos, B. (2015). 'Goal-oriented performance evaluation methodology for page segmentation techniques', 13th International Conference on Document Analysis and Recognition (ICDAR), 281-285.

Konstantinos, Z., Pratikakis, I., Antonacopoulos, A., Gatos, B., and Papamarkos, N. (2012). 'Handwritten and Machine Printed Text Separation in Document Images Using the Bag of Visual Words Paradigm", in: Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference, Bari, 103-108. DOI: 10.1109/ICFHR.2012.207.

Introduction to electronic books: EPub 3.0 and beyond

C.M. Sperberg-McQueen
cmsmcq@acm.org
Black Mesa Technologies LLC, United States of America

Liam Quin
liam@w3.org
World Wide Web Consortium (W3C), Canada

A short introduction to current standards for electronic books, focusing on EPub 3.0 (a standard issued by the International Digital Publishing Federation) and its supporting specifications: XHTML, HTML 5, CSS, SVG, and Zip.

Are the dreams or visions of Ted Nelson, Alan Kay, and other visionaries of electronic books finally coming true? Current tablets, smart phones, and dedicated ebook devices seem to bring them closer to daily reality than ever before. We will survey the current shape of the territory for ebooks: what they can do, what they currently cannot do, what we hope they will be able to do soon. We will demonstrate some of the software available for generating EPubs and for checking them for device- and software-specific problems. Finally, we will outline some outstanding design and production challenges.

Many of the challenges of ebook production will be familiar to anyone with experience in book or journal production: mathematics, tables, graphics, and figures are no easier to handle (but, happily, not always much harder) in electronic books than they are in print publications or on the Web. Others will be familiar from web-site production: ebook readers and applications vary both in terms of what then can do and in terms of how closely they adhere to the EPub specification. The requirement that ebooks be self-contained and work with no network connection brings new challenges, as does the the need for accessible content. Participants will learn how to check against relevant specifications, how to research problems, and how to attain an intuition for the sorts of things that are both possible and practical today.

The International Digital Publishing Federation (IDPF), which developed EPub, has recently merged with the World Wide Web Consortium; we will discuss the outlook for further work in the area and explain how to get involved and how to help ebook readers improve to meet your needs.

Prerequisites: no firm prerequisites. Participants with some familiarity with XML, HTML, HTML 5, and CSS will be in a better position to follow the details of examples.

XQuery for Digital Humanists

Joseph Wicentowski
joewiz@gmail.com
Office of the Historian, U.S. Department of State
United States of America

Clifford Anderson
clifford.anderson@vanderbilt.edu
Vanderbilt University, United States of America

Introduction

This half-day tutorial introduces digital humanists at any level of experience to XQuery, a mature, high-level programming language used in many DH projects because it is purpose-built for analyzing, manipulating, and publishing data stored in the XML-based data formats that many DH projects use, e.g., TEI, EAD, MODS, METS. Prominent XQuery-based projects include Carl Maria von Weber Gesamtausgabe, Foreign Relations of the United States and Syriaca.org.

Led by two experts who each have a decade of experience using and teaching XQuery and who have co-authored *XQuery for Humanists* (forthcoming, Texas A&M University Press), this half-day tutorial introduces the key concepts underlying the XQuery language and the kinds of analysis that it makes possible. The focus will be on exploring TEI-encoded editions with simple XQuery expressions.

Using a free and easy to install XQuery learning environment, participants (who must bring their own laptops) will gain hands-on experience writing queries against open datasets, including a TEI-encoded documentary edition, *Foreign Relations of the United States*. Participants will gain a basic foundation in the language and be introduced to community resources for further study.

This half-day workshop will cover the basics of XQuery, providing participants with sufficient hands-on experience to start exploring their own scholarly editions and metadata with XQuery. We presuppose that participants will have come with a basic understanding of XML and TEI.

Brief outline

I. Introduction: XQuery for the Digital Humanities
II. Setting up an XQuery environment
III. Finding data with XPath
IV. Writing FLWOR expressions
V. Exploring XQuery Full-Text

Each section (except the introduction) will include hands-on exercises.

Target audience:

Students, scholars, and practitioners who use or are interested in using digital methods in their humanities work in academic departments, libraries, "alt-ac" fields, or their private capacity; no previous programming experience required; some experience XML or an XML-based format (TEI, EAD, MODS, METS) useful but not required. Participants will work with a common dataset provided by the tutorial leaders, but they may bring their own datasets for practice during the lab and consultation period.

Tutorial leaders

Clifford B. Anderson is Associate University Librarian for Research and Learning at Vanderbilt University in Nashville, Tennessee. He has a M.Div. from Harvard Divinity School and a Th.M. and Ph.D. from Princeton Theological Seminary. He also holds a M.S. in Library and Information Science from the Pratt Institute in New York City. Cliff started working with XQuery in 2006 before the first official version of the language was released. In 2014, he served as the project leader of the NEH-funded XQuery Summer Institute at Vanderbilt University. He has also taught sessions on XQuery for iterations of Laura Mandel's Programming for Humanists course at Texas A&M and leads the weekly XQuery working group at Vanderbilt University for digital humanists

Joseph C. Wicentowski is the Digital History Advisor in the Office of the Historian at the U.S. Department of State. He received his Ph.D. from Harvard University in modern East Asian history. He started using XQuery in 2007 to analyze and publish the Office of the Historian's TEI-encoded publications and datasets. For more on the project, see Wicentowski (2011). All code and data from the project are freely available on GitHub. He recognized XQuery's potential to empower students, scholars, and practitioners to take control of their own data and build their own applications. But he knew that without resources geared toward people with a humanities background, others would struggle as he first did. He began writing about XQuery in various digital humanities forums, contributing to the XQuery Wikibook online textbook, and giving workshops at the TEI@Oxford and Digital Humanities@Oxford Summer School programs. Joe regularly speaks and writes in the fields of history, documentary editing, and open government. He also actively participates in TEI, XQuery, and digital humanities communities, and fosters discussion about XQuery on Twitter at @XQuery.

Bibliography

Wicentowski, J. (2011) "history.state.gov: A case study of Digital Humanities in Government," *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, vol. 1 no. 3 , https://letterpress.uchicago.edu/index.php/jdhcs/article/view/80.

XQuery for Data Integration

Joseph Wicentowski
joewiz@gmail.com
Office of the Historian, U.S. Department of State
United States of America

Clifford Anderson
clifford.anderson@vanderbilt.edu
Vanderbilt University, United States of America

Introduction

This half-day tutorial shows how XQuery integrates digital humanities data from multiple sources and formats. Drawing on the latest features of XQuery 3.1, the instructors demonstrate how to draw together information from the most common structured data formats, namely, JSON, CSV, RDF, and XML. We will teach some of the latest features of the XQuery language, including how to work with maps, arrays, and new functions like json-doc(), parse-json() and json-to-xml().

Specifically, we will explore the following data sources: (1) dictionary data (in JSON) from the Oxford English Dictionary; (2) an Open Publication Distribution System (OPDS)-based ebook catalog that makes publications at the U.S. Department of State searchable, browse-able and downloadable via OPDS-compliant ebook reader apps like Shubook, Hyphen, etc.; (3) an OpenRefine reconciliation endpoint API built to let people run their own lists of people against biographical databases; and (4) interacting with IIIF APIs.

Using a free and easy to install XQuery learning environment, participants (who must bring their own laptops) will gain hands-on experience writing queries against open datasets in CSV, JSON, and RDF and integrating this data with XML. Participants will gain exposure to the latest features of the XQuery language as well as best practices for connecting data across systems and formats. We presuppose that participants will have come with a basic understanding of XQuery.

Brief outline

I. Requesting remote data and storing it into an XML database
II. Querying CSV with XQuery
III. Querying JSON with XQuery
IV. Querying RDF with XQuery
V. Enriching TEI with data from other sources and formats

Each section will include hands-on exercises.

Target audience

Students, scholars, and practitioners who use or are interested in using digital methods in their humanities work in academic departments, libraries, "alt-ac" fields, or their private capacity; no previous programming experience required; some experience with XML or an XML-based format (TEI, EAD, MODS, METS, Atom) useful but not required. Participants will work with a common dataset provided by the tutorial leaders, but they may bring their own datasets for practice during the lab and consultation period.

Tutorial leaders

Clifford B. Anderson is Associate University Librarian for Research and Learning at Vanderbilt University in Nashville, Tennessee. He has a M.Div. from Harvard Divinity School and a Th.M. and Ph.D. from Princeton Theological Seminary. He also holds a M.S. in Library and Information Science from the Pratt Institute in New York City. Cliff started working with XQuery in 2006 before the first official version of the language was released. In 2014, he served as the project leader of the NEH-funded XQuery Summer Institute at Vanderbilt University. He has also taught sessions on XQuery for iterations of Laura Mandel's Programming for Humanists course at Texas A&M and leads the weekly XQuery working group at Vanderbilt University for digital humanists

Joseph C. Wicentowski is the Digital History Advisor in the Office of the Historian at the U.S. Department of State. He received his Ph.D. from Harvard University in modern East Asian history. He started using XQuery in 2007 to analyze and publish the Office of the Historian's TEI-encoded publications and datasets. For more on the project, see Wicentowski (2011). All code and data from the project are freely available on GitHub. He recognized XQuery's potential to empower students, scholars, and practitioners to take control of their own data and build their own applications. But he knew that without resources geared toward people with a humanities background, others would struggle as he first did. He began writing about XQuery in various digital humanities forums, contributing to the XQuery Wikibook online textbook, and giving workshops at the TEI@Oxford and Digital Humanities@Oxford Summer School programs. Joe regularly speaks and writes in the fields of history, documentary editing, and open government. He also actively participates in TEI, XQuery, and digital humanities communities, and fosters discussion about XQuery on Twitter at @XQuery.

Bibliography

Wicentowski, J. (2011) "history.state.gov: A case study of Digital Humanities in Government," *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, vol. 1 no. 3 , https://letterpress.uchicago.edu/index.php/jdhcs/article/view/80.

Computer Vision in Digital Humanities

Martijn Kleppe
martijn.kleppe@kb.nl
National Library of the Netherlands, the Netherlands

Matthew Lincoln
mlincoln@getty.edu
Getty Research Institute, United States of America

Melvin Wevers
m.j.h.f.wevers@uu.nl
Utrecht University, the Netherlands

Mark Williams
mark.j.williams@dartmouth.edu
Dartmouth College, United States of America

Benoit Seguin
benoit.seguin@epfl.ch
École Polytechnique Fédérale de Lausanne, Switzerland

Thomas Smits
t.smits@let.ru.nl
Raboud University, the Netherlands

Introduction

Although the majority of Digital Humanities scholars still focus on textual analysis, we see an increasing number of studies using digitised visual sources and taking the first steps in the field of 'Visual Big Data' (Ordelman et al, 2014). Scholars have increasingly used both large-scale digitised visual datasets as well as computational methods to analyse these datasets in new ways, see for example the work of Lev Manovich on Manga comics (2012), Time covers, and selfies (Manovich & Tifentale 2015) or the recent work of Lenardo et al (2016) on visual pattern discovery in large databases of paintings as presented at the DH 2016 conference. Others focus on the identification of finding recurring images in visual collections. Terras & Kirton (2013) analyse the reuse of digital images of cultural and heritage material on the internet while Resig (2014) and Reside (2015) follow a similar approach using a closed dataset.

A similar trend can be observed at the ADHO Digital Humanities Conferences. In 2014 the special Interest group Audiovisual Materials in Digital Humanities was founded and at the 2015 and 2016 conferences, several papers were presented using (audio) visual sources (see e.g. Kleppe 2015, Lenardo et al. 2016, Lincoln 2016). Based on an analysis of the submissions for the 2017 conference, Scott Weingart observes that an increasing number of papers are based on research concerning non-textual sources (Weingart 2017). The use of non-textual sources challenges established research practices of Digital Humanities, leading to new research questions and the need for transformation of existing approaches, and the development of new methodologies. For example, how can we analyse the characteristics of images in large visual datasets? Some of these questions have also been raised in the analogue era. See, for example, the work of Barry Salt (1974) on the characteristics of opening shots of twentieth-century films, Scott McCloud (1994) on the visual language of Japanese manga and comics, or Peter Burke's take on using visual materials as historical evidence (2001).

Due to the digital turn, we now see the application of new techniques to answer similar and new types of research questions. One of the dominant techniques to analyse large-scale visual datasets is computer vision: a field that deals with how computers can generate a high-level of understanding of visual material (Smeulders 2000).This field has been seeing unprecedented improvements as part of the "Deep Learning Revolution" since the first deep Convolutional Neural Network (Krizhevsky 2012), unleashing new possibilities. This workshop will focus on how computer vision can be applied within the realm of Audiovisual Materials in Digital Humanities. During the workshop, attendees will both present (ongoing) work on applying computer vision and experiment with computer vision in their own work in a hands-on session. Possible questions include:

- What are state-of-art techniques in computer vision?
- What are possibilities and challenges in the application of computer vision in Digital Humanities research?
- How can non-expert researchers employ computer vision?

- How can cooperation between Humanities researchers and computer vision experts be improved so both can benefit of each other's expertise?

Workshop outline

The workshop will consist of four pillars:

- **A keynote** by Lindsay King (Associate Director for Access and Research Services Robert B. Haas Family Arts Library, Yale University) & Peter Leonard (Director Digital Humanities Lab, Yale University Library) & on "Processing Pixels: Towards Visual Culture Computation"
- **Paper presentations** with results and ongoing work on applying computer vision in DH research projects. Papers will be selected by a review commission. All details on the Call for Abstracts can be found at the workshop website.
- **A hands-on session** in which participants will be able to experiment with open source Computer Vision tools. This session will be led by Benoit Seguin of École Polytechnique Fédérale de Lausanne, (EPFL). Benoit Seguin is a PhD Student in Computer Science from the DHLAB at EPFL. His work focuses on automatic patterns detection across iconographic collections. Before starting his PhD, he worked with microscope images for biomedical applications or electronic manufacturing problems.
- **Lightning Talks** allowing participants to share their ideas, projects or ongoing work in a short presentation of two minutes. Interested researchers who want to present work will be able to submit their idea in a later stage.

SIG Endorsement

This workshop proposal is endorsed by the steering committee of the Special Interest Group Audiovisual Materials in the Digital Humanities (AVinDH). This will be the third workshop that will be organised by the SIG. The first workshop on audiovisual material in digital humanities took place at the 2014 DH Conference in Lausanne. This workshop paved the way for the foundation of the AVinDH Special Interest Group that was officially launched during the DH2015 Conference in Sydney. The second workshop took place during the DH2016 Conference in Cracow and dealt with audiovisual material and dedicated analysis tools in DH.

While the previous SIG-endorsed workshops were organised by the SIG's steering committee, this proposal on Computer Vision in Digital Humanities is mainly organised by members of the SIG and one representative of the Steering Committee. The steering committee applauds the initiative of the organisers since the theme of the workshop fits perfect in both the SIG's aims as well as the current interest in non-textual sources as observed by Scott Weingart in his first analyses of the DH2017 submissions. Furthermore, the steering committee is enthusiastic about providing a platform for members of the SIG and the complete DH community to present high-quality work that will be selected by a review committee as well as ongoing work that will be presented in lighting talks. Both formats will contribute to the community and network building function that the SIG AVinDH promotes.

Programme Committee

The programme committee (PC) for this workshop consists of the workshop proposers (see list with names and bios below). For the paper selection process, the PC will be extended with at least 10 experts. The selection process will be chaired by Prof. Franciska de Jong, Utrecht University/CLARIN ERIC and chair of the SIG's Steering Committee (f.m.g.dejong@uu.nl)

Composition of the Programme Committee (workshop proposers only)

Dr. Martijn Kleppe

Martijn Kleppe is a researcher at the Research Department of the National Library of the Netherlands (KB) where he works on several Digital Humanities Projects as part of the KB's Digital Humanities Team. Before he worked as academic researcher at the Erasmus University Rotterdam and Vrije Universiteit Amsterdam. He wrote his dissertation on 'Canonical Iconic Photographs' and was involved in several Digital Humanities projects focussing on opening up (audio)visual archives.

Melvin Wevers

Melvin Wevers is in the final stage of his PhD research at Utrecht University on the role of the United States as a reference culture in twentieth-century Dutch newspaper debates on consumer goods. In his research, he primarily focuses on text analysis but he is moving in the direction of computational analyses of visual materials. He also worked on the eScience project *ShiCo: Shifting Concepts Over Time* in which he used word embeddings to study conceptual change in large historical corpora. In Spring 2016, he was a research fellow of the 'culture analytics' program hosted by the UCLA's Institute of Pure and Applied Mathematics. In March 2017, he will be researcher-in-residence at the National Library of the Netherlands (KB) to work on the computational analyses of images in newspaper advertisement.

Mark Williams

Mark Williams is Associate Professor of Film and Media Studies at Dartmouth College. He received both of his graduate degrees in Critical Studies from The School of Cinema-Television at The University of Southern California. He has published in a variety of journals and anthologies, including Télévision: le moment expérimental (1935-1955); Convergence Media History; New Media: Theories and Practices of Digitextuality; Collecting Visible Evidence; Dietrich Icon; Television, History, and American Culture: Feminist Critical Essays; and In Living Color: Race, Feminism, and Television. He directed the Leslie Center Humanities Institute entitled Cyber-Disciplinarity. In conjunction with the Dartmouth College Library, he is the founding editor of an e-journal, The Journal of e-Media Studies. With Adrian Randolph, he co-edits the book series Interfaces: Studies in Visual Culture for the University Press of New England. With Michael Casey, he received an NEH Digital Humanities Start-Up Grant to build the ACTION toolset for cinema analysis. He received an award for Scholarly Innovation and Advancement at Dartmouth for directing The Media Ecology Project (MEP). He has published about MEP in The Arclight Guidebook to Media History and The Digital Humanities (2016), and also in The Moving Image (2016). Most recently Williams has received with John Bell (MEP architect) an NEH grant to build a Semantic Annotation Tool (SAT). With Lorenzo Torresani (Dartmouth Computer Studies) he has received a Knight Foundation grant to develop computer vision and machine learning capacities for moving images.

Thomas Smits

Thomas Smits is completing a PhD on the transnational trade in illustrations of the news and the production of identity by mid-nineteenth century European illustrated newspapers at the Radboud University in the Netherlands. He is an editor of the *Journal for European Periodicals Research* (JEPS) and a PhD-board member of the Royal Netherlands Historical Society. In May 2017 he will start a researcher-in-residence project at the National Library of the Netherlands (KB) entitled *Illustrations to Photographs: using computer vision to analyse news pictures in Dutch newspapers, 1860-1940.*

Benoit Seguin

Benoit Seguin is a PhD student at the Digital Humanities Laboratory of EPFL. His work focuses on using modern computer vision algorithms to navigate large iconographic collections, and he recently showed how Neural Networks could be trained to recognize similar patterns in different artworks. He is part of the bigger REPLICA project which aims at digitizing and making searchable the 1M photos of the Cini collection in Venice. He took part of the organization of the recent "New Methods and Technologies for Art History" Summer School between ETHZ and EPFL. Before starting his PhD, he got his MSc in Computer Science with a thesis about SEM Image Analysis from IBM Research Zurich.

Matthew Lincoln

Matthew Lincoln (Ph.D. University of Maryland 2016) is a Data Research Specialist with the Getty Research Institute, working on the Provenance Index Remodeling Project. He specializes in computationally-driven analysis of art history, including complex network analysis and machine learning methods on cultural datasets. He has previously worked as a curatorial fellow with the National Gallery of Art in Washington, DC, and has received grants from both the Getty and Kress Foundations as part of digital art history summer institutes, and in 2016 was an organizer for a jointly-funded conference "Digital Dimensions of Art History" hosted by the University of Maryland and the Maryland Institute for Technology and Humanities. He has published data-driven research articles in the *International Journal of Digital Art History* and *British Art Studies*, and is a contributor to *The Programming Historian*.

Target audience

This workshop aims to bring together scholars from different fields who have an interest in or actively use computer vision to analyse (large) datasets of digitised audiovisual sources. We have a wide network of peers that can be reached, and will also scan the papers of the conference that have been selected with the aim to invite potential contributors in this realm to present their work in our workshop. We aim at participation of 20 – 30 persons.

Bibliography

Burke, P. (2001). Eyewitnessing. *The uses of images as historical evidence*. London: Cornell University Press

Kleppe, M. (2015) Tracing the afterlife of iconic photographs using IPTC. Digital Humanities 2015, 29 Juni - 3 Juli 2015, Sydney

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet classification with deep convolutionnal neural network

di Lenardo, I., Seguin, B., Kaplan, F. (2016). Visual Patterns Discovery in Large Databases of Paintings. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 169-172. http://dh2016.adho.org/abstracts/348

Lincoln, M. (2016). If Paintings were Plants: Measuring Genre Diversity in Seventeenth-Century Dutch Painting and Printmaking. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 256-259. http://dh2016.adho.org/abstracts/133

Manovich, L. (2012). How to compare one million Images? In Berry, D. M., *Understanding Digital Humanities*, pp. 249-78.

Manovich. L. and Tifentale, A. (2015). Selfiecity: Exploring Photography and Self-Fashioning in *Social Media*. In Berry, David M., Dieter, M. (eds), Postdigital Aesthetics: Art, Computation and Design, pp. 109-22.

McCloud, S . (1994). *Understanding Comics: The Invisible Art.* New York: HarperPerenn

Ordelman, R., Kemman, M., Kleppe, M., de Jong, F., Scagliola, S. (2014) Sound and (moving) Images in Focus - How to integrate Audiovisual Material in Digital Humanities Research. *Digital Humanities 2014* http://dharchive.org/paper/DH2014/Workshops-914.xml

Reser, G. and Bauman, J. (2012). The Past, Present, and Future of Embedded Metadata for the Long-Term Maintenance of and Access to Digital Image Files. *International Journal of Digital Library Systems (IJDLS)*, 3(1): 53-64.

Reside, D. (2014). Using Computer Vision to Improve Image Metadata. *Digital Humanities 2014*. http://dharchive.org/paper/DH2014/Paper-294.xml

Salt, B. (1974). The Statistical Style Analysis of Motion Pictures. *Film Quarterly*, 28(1): 13-22.

Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. Pattern Analysis and Machine Intelligence, *IEEE Transactions o*n Pattern Analysis and Machine Intelligence, 22(12), 1349-138

Weingart, S. (2017), *Submission to DH2017 (pt. 1)* http://scottbot.net/submissions-to-dh2017-pt-1/

Terras, M. M. and Kirton, I. (2013). Where do images of art go once they go online? A Reverse Image Lookup study to assess the dissemination of digitized cultural heritage. *Selected papers from Museums and the Web North America*, pp. 237–48. http://mw2013.museumsandtheweb.com/paper/where-do-images-of-art-go-once-they-go-online-a-reverse-image-lookup-study-to-assess-the-dissemination-of-digitized-cultural-heritage/

Shaping Humanities Data: Use, Reuse, and Paths Toward Computationally Amenable Cultural Heritage Collections

Thomas Padilla
thomaspadilla@ucsb.edu
UC Santa Barbara, United States of America

Sarah Potvin
spotvin@library.tamu.edu
Texas A&M University, United States of America

Laurie Allen
laallen@upenn.edu
University of Pennsylvania, United States of America

Stewart Varner
svarner@sas.upenn.edu
University of Pennsylvania, United States of America

Galleries, libraries, archives, and museums (GLAMs) increasingly seek to make digitized and born-digital collections accessible as data optimized for computational methods and tools common to the Digital Humanities. Preparation and publication of collections as data extends possible collection use beyond the analog object interactions that collection interfaces tend to try and emulate. In line with open data efforts, libraries, archives, and museums typically work to assign open licenses to these data. Current access methods are widely divergent, spanning simple provision of compressed collection objects in ZIP files, exposing static collection websites that can be crawled using a tool like rsync, leveraging Github for text collection access, provisioning an API, enabling FTP access to collections, mediating computational processes performed on collection data through a platform, to facilitating data access through use of torrent technology. Concurrently, in response to researcher requests for data-mining, commercial publishers have developed a range of processes for delivering proprietary corpuses with terms and conditions that significantly limit or expressly forbid data sharing, including providing libraries with physical hard drives loaded with the data. There are no consensus-driven best practices that guide the generation, description, and provisioning of computationally amenable GLAM collections for the range of communities that fall within the Digital Humanities. Without best practices in this space, institutions run the risk of misplaced investment of resources that foster the creation of irregular, ultimately disorienting data access environments. Indeed, the panoply of institutional approaches poses a challenge to GLAM institutions seeking best practices and clear guidelines for publishing collections as data.

One major barrier to the development of consensus-driven best practice is an incomplete understanding of *how* digital humanists, among others, are using and reusing cultural heritage data. This workshop aims to make progress towards bridging that gap. Research indicates that types of use exhibited by digital humanists include but are not limited to text analysis, image analysis, mapping, sound analysis, and network analysis. Orientation to the full scope of *academic* use types can be gained through in-depth analysis of data use practices across disciplines as represented in core Digital Humanities journals (Padilla and Higgins 2016), by reviewing works at the annual global Digital Humanities conference (Weingart 2016), and by studying edited volumes that have to this point effectively compiled a broad range of research in this space (Gold 2012; Gold and Klein 2016; Burdick, Drucker, Lunenfeld et al 2012; Schreibman, Siemens, Unsworth 2016).

This workshop will build upon this orientation by engaging directly with digital humanists' existing and projected research and pedagogical practices that draw upon ever growing GLAM collections. Blending short talks by practitioners, guided discussion, and workshopping of the organizers' draft framework (further described below), the workshop will focus on how researchers and educators use GLAM collections that have been made accessible as data, and will extend to consider how these uses should inform collection creation and access.

The organizers of this workshop are members of the project team for "Always Already Computational: Library Collections as Data," an effort sponsored by the Institute of Museum and Library Services in the United States of America through their National Forum grant program. The organizers have observed that GLAM approaches to the preparation of collections as data are often heavily influenced by national or regional priorities and associated infrastructures. Yet the use and reuse of these open data is necessarily international. While the organizers of the workshop are US-based, the workshop aims to surface geographically-diverse praxis. The short talks in the workshop have been selected through an open CFP facilitated by an international program committee.

The workshop may be structured thematically, based on talks and demos solicited via the CFP. Participants will be encouraged to consider how efforts to develop computationally amenable collections, which run the risk of recreating and reinforcing long standing biases inherent in cultural heritage collection practice, provide an opportunity to reframe, enrich, and/or contextualize collections in a manner that seeks to avoid replication of bias.

Potential themes may include:

- **Use and Reuse:** How are data used and reused? What methods and tools are commonly employed? Do these differ by disciplinary community? What types of data are used? What types of data are desired but are difficult to use for reasons included but not limited to copyright status, content type (e.g. video, audio, web, software), size? How, when, and where are data reused? What factors enhance or inhibit the likelihood of data reuse?
- **Access:** What can we learn from our collective experiences working to access data from within and outside of the cultural heritage community? What are preferred methods of data access? What factors should be considered when deciding among access methods? When is simple click and download of bulk collections appropriate? What characteristics define an optimally useful API (application programming interface) for a wide range of users with varying technical expertise? Is an API always the best route to go? Are there a mix of options that should be considered? What considerations inform development of those access options?
- **Description and Discovery:** How do digital humanists locate appropriate data? What tools are used to search for data? What information about the data is necessary to enable use? When compiling meta-collections of data, how are digital humanists maintaining provenance and merging disparate metadata?

Bibliography

Burdick, A., Drucker, J. and Lunenfeld, P., Presner, T. and Schnapp, J. (Eds) (2012). *Digital_Humanities*. Cambridge: MIT Press. https://mitpress.mit.edu/books/digitalhumanities

Gold, M. (Ed) (2012). *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/1

Gold, M. and Klein, L. (Eds) (2016). *Debates in the Digital Humanities.* Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/2

Padilla, T., Higgins, D. (2016). Data Praxis in the Digital Humanities: Use, Production, Access. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 644-646.

Schreibman, S., Siemens, R. and Unsworth, J. (Eds). (2016). *A New Companion to Digital Humanities*. 2nd edition. Wiley-Blackwell.

Weingart, S. (2016). “Submissions to DH2016 (pt. 1).” On *the scottbot irregular*. http://scottbot.net/submissions-to-dh2016-pt-1/

DH 2017