Building the Semantic Portal of Italian Divagrafie

Laura Pandolfo*, Lucia Cardone, Luisa Cutzu, Beatrice Seligardi and Giulia Simi

DUMAS, University of Sassari, via Roma 151, Sassari, Italy

Abstract

In this paper, we present the preliminary research activities on building a semantic digital archive for publishing heterogeneous data about a literary corpus provided by Italian actresses, known as Divagrafie. This corpus represents a unique collection of cultural data for scholars in Film and Literature Studies through which they can analyze the phenomenology, characteristics, and historical evolution of the writings produced by Italian actresses. In this paper, we present the vision behind the development of the semantic digital archive and explore the potential applications and its expected impacts in the research community and society.

Keywords

Semantic Web, Digital Archive, Film and Literature Studies

1. Context and Motivation

The availability of cultural heritage data has rapidly boosted the emerging new research area in Digital Humanities (DH) [1] where an increasing number of scholars are dealing with new computational methods developed and applied to literary corpus for solving problems in humanities and social sciences. Recently, DH research has shifted the focus to providing the user with integrated tools for solving research problems in interactive ways [2]. One of the characteristics of this kind of systems is to use computational methods for solving humanities research questions by using large datasets and applying on them the so-called "distant reading" approach [3], namely a set of different techniques, such as novel statistical methods, sentiment analysis, topic modeling and network analysis. The main benefit of using the distant reading approach comes from the fact that it opens new horizons for computational literary studies, without the hard effort required by classical humanistic computer research, which often needs rigorous coding and document annotations. However, this approach has several critical issues, as identified by Ciotti [4, 5]. One shortcoming lies in the fact that most of the applied computational methods are independent from the context, while humanities and literary data need to be heavily contextualized. Another important critical point concerns the interpretation of text, which is usually an intentional process. Statistical computational methods are hardly able to detect the

IPS-RiCeRcA-SPIRIT 2022: 10th Italian Workshop on Planning and Scheduling, RiCeRcA Italian Workshop, and SPIRIT Workshop on Strategies, Prediction, Interaction, and Reasoning in Italy.

*Corresponding author.

lpandolfo@uniss.it (L. Pandolfo); cardone@uniss.it (L. Cardone); lcutzu@uniss.it (L. Cutzu); bseligardi@uniss.it (B. Seligardi); gsimi@uniss.it (G. Simi)

0000-0002-5785-5638 (L. Pandolfo); 0000-0003-0770-6324 (L. Cardone); 0000-0002-9400-0866 (B. Seligardi); 0000-0002-2206-2945 (G. Simi)

true interpretation since the meaning of a word is usually determined by the attribution of sense and meaning by the author and by the reader.

Semantic Web (SW) [6] technologies and Linked Data [7] can overcome these issues by enriching the distant reading approach of new methods able to capture the semantic nature of literary texts [8]. Therefore, cultural heritage has become an active area of application of SW, where cultural content and metadata are available openly for research and public use based on collections in museums, libraries, archives, and media organizations [2, 9]. In the last two decades, large amount of data has been aggregated in huge national and international portals, libraries and repositories such as Europeana [1] by forming a significant part of DBpedia [2] and Wikidata [3]. The concept of ontology [10] plays a central role, since it is commonly used as a sort of schema capturing knowledge about a specific domain via providing relevant concepts and relations between them. Currently,several examples of ontology-based digital archives and libraries in the humanities have been reported [11, 12, 13, 14, 15, 16].

This paper introduces our ongoing research work on developing an ontology-based archive named *WOmen Writing around the camera* (WOW) which will collect semantic data about relevant writings produced by Italian actresses ("Divagrafie") focusing on the dynamics of exchange between writing, acting performance and the construction of the star image. This research idea builds-on and extends the "Drawing a Map of Italian Actresses in Writing" (DaMA) funded by PRIN 2017 [4] which aims at investigating the extension, phenomenology, characteristics, and historical evolution of the writings produced by Italian actresses. The main goal is to build a semantic portal containing different resource materials related to Italian actresses' visual and self-representation history able to bring to light and interrogating together different kinds of documents (mainly photographs and texts) with a defined set of investigation methods with the help of DH tools, particularly those based on SW technologies and distant reading methods. The paper is organized as follows. Section 2 describes our current research progress and outlines the planned phases related to the development of the WOW semantic archive, while Section 3 concludes the paper by discussing the potential applications and expected impacts resulting from this line of research both for the academic point of view but also society in general.

2. The WOW Semantic Archive

The idea to develop the WOW Semantic Archive stems from the in-progress research project DaMA which aims at investigating the extension, phenomenology, characteristics, and historical evolution of the writings produced by Italian actresses, focusing in particular on the dynamics of exchange between writing, acting performance and construction of the star image. Defined as Divagrafie [17], such writings represent a multifaceted and interesting corpus of texts traditionally overlooked by the academic community. The important contribution to the understanding of stardom provided by this kind of self-narratives or fiction writings has been only partially acknowledged within the field of star studies; similarly, the field is also lacking

[1] https://www.europeana.eu/en

[2] https://www.dbpedia.org/

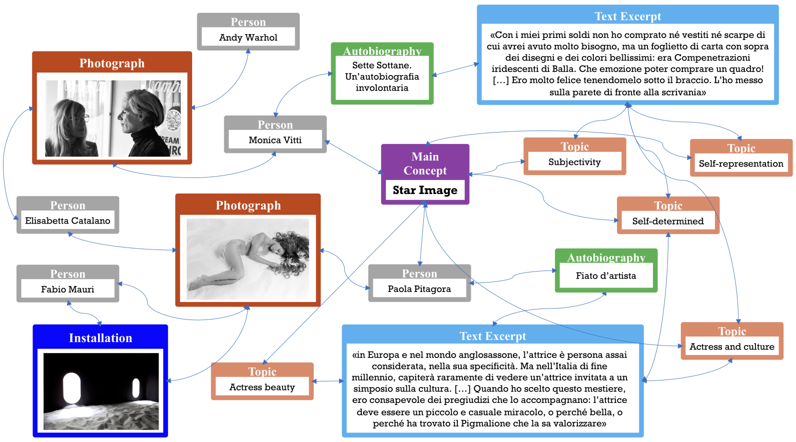
[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

[4] https://www.damadivagrafie.org/

an in-depth analysis dedicated to the relationship between the writings of a specific actress and the construction of her own star image. This ongoing project is revealing a still unknown territory and has so far classified a wide number of texts, including 80 autobiographies written by 47 actresses. This literary corpus – which we are intended to focus on – represents a unique testing ground for a convergent methodology that applies DH methods, SW technologies and the most recent advances in the fields of Film and Media Studies, Literature and Gender Studies. These volumes have been analyzed so far by the DaMA team with a *close reading* approach focusing on paradigmatic examples and eliciting a variety of recurring topics [18, 19, 20], with a particular focus on the interconnections with the perspective of performing and actors' studies, stardom and celebrities studies and by using also the videographic analysis tools.

By following and expanding the multidisciplinary approach already explored by the DaMA research, our current and future research work focuses on increasing the actual corpus and applying on it SW standards and DH tools. Below the main phases of our methodological approach:

1. **Recognition of DaMA project's outcomes and integration of new materials**. All the materials collected and analyzed by the DAMA research project team will be subjected to a further and extended analysis in order to provide a first taxonomy describing the corpus. During this phase, new documents and resources will be integrated into the original corpus, in particular we we will focus on the retrieval of materials preserved in the Elisabetta Catalano archive (Rome), the Museo di Fotografia Contemporanea - MuFoCo archive (Cinisello Balsamo, MI), the Museo del Cinema archive (Turin), the Cinemazero archive, the Luisa Di Gaetano archive (Rome), Archivia - Casa Internazionale delle Donne (Rome). In addition, we intend to study cinema and cultural periodicals in order to verify the circulation and the incidence of the work of female photographers in the national press.
2. **Conceptualization and formalization**. The main goal of this activity is to formalize the acquired knowledge in terms of concepts and relations with respect to specific ontological schema and models detected in this phase. In this domain, the use of ontologies provides a range of benefits for the users, e.g., in searching and browsing by concept rather than string-based only. For example, the photographs that Elisabetta Catalano has taken of a variety of actors and actresses, writers and directors can be put in relationships with all the excerpts in the Italian actresses' autobiographies in which the relations with the cultural field strongly emerge as a symptom of overcoming the stereotype of actresses as a pure "not thinking" beautiful and fashioned body. As we can see in Figure 1, Elisabetta Catalano's photograph of Monica Vitti talking to Andy Warhol can be intertwined with passages in her autobiographies related to art as a means of self-expression and with other several entities and concepts in the corpus.
3. **Semantic annotations and analysis of the corpus**. In this phase, ontological schema and models will be used to guide the semantic annotations of all the corpus materials (texts and images). During this phase, the corpus will be enriched with metadata describing, for example, references that link the content to specific concepts. Moreover, a variety of literary analysis based on distant reading approaches will be applied to the corpus materials. In particular, we will explore typical computational linguistics methods based



«Con i miei primi soldi non ho comprato né vestiti né scarpe di cui avrei avuto molto bisogno, ma un foglietto di carta con sopra dei disegni e dei colori bellissimi: era Compenetrazioni iridescenti di Balla. Che emozione poter comprare un quadro! […] Ero molto felice tenendomelo sotto il braccio. L'ho messo sulla parete di fronte alla scrivania»

«in Europa e nel mondo anglosassone, l'attrice è persona assai considerata, nella sua specificità. Ma nell'Italia di fine millennio, capiterà raramente di vedere un'attrice invitata a un simposio sulla cultura. […] Quando ho scelto questo mestiere, ero consapevole dei pregiudizi che lo accompagnano: l'attrice deve essere un piccolo e casuale miracolo, o perché bella, o perché ha trovato il Pigmalione che la sa valorizzare»

Figure 1: Example of entities and relationships in the examined corpus.

on Natural Language Processing techniques in order to identify recursive narrative patterns as well as similar characteristics of the texts which will allow the recognition of an attantial typology specifically linked to Divagrafie.

4. **Development of the semantic archive**. The main goal of this last phase will be the development of the WOW semantic archive. It will provide the researchers with a set of powerful and efficient tools that can be used to query, analyze and study the data in order to make possible intertwine and connect different data. The interdisciplinary approach will be made visible and easily traceable by a digital user interface able to link different objects and produce new sets of knowledge. This will be particularly useful for scholars and researchers of different fields, such as History of Cinema, History of Photography, Visual Studies, Literature, Contemporary History and so on. The semantic archive will be the hub through which not only find materials to be linked to other studies and researches, but also to find new ways to analyze and study them with innovative approaches. In this phase, we intend to investigate some automatic techniques for ontology population, such as those presented in [21, 22]. All the features of the semantic archive will be integrated into an easy-to-use graphical interface which will provide the visualization of the data in various formats, such as graphs, interactive maps, timelines, facets, etc.

3. Potential Applications and Expected Impacts

The development of the WOW Semantic Archive could have significant potential applications within the national and international research communities involved in the process. In fact, the archive will make accessible a wide corpus of materials related to Italian Actresses which crosses a wide range of studies also in the international academic communities: Film Studies, Photography Studies, Women's Studies and Literature Studies, and so on. To our knowledge, this is the first time that such a corpus is being created following the unambiguous, rigorous, consistent and well-documented practices provided by the DH approach. Moreover, this is the first time that computational methods and SW technologies will be applied to this kind of document resources. The data integration feature provided by SW technologies will allow the connection between data contained in the WOW Semantic Archive to data included to other international archives, such as the European Film Gateway [5], by increasing the international impact.

The semantic archive will be useful also for upper-secondary teachers and students, curators of film-related events, film/photography/literature enthusiasts, and in general for whoever needs to retrieve the data and the researches included in it. In addition to the ambition to fill a gap in research and reflection on the area of study, namely the analysis of the relationship between women, photography and cinema from a feminist perspective, which in many Western countries has already been addressed, this work has another ambition, more complex and therefore even more important and stimulating: this research work aims at contributing to the debate and to the cultural framework of Italy, in which gender issues, women's emancipation, feminist heritage, women's art are still not sufficiently considered as a fundamental part of the cultural heritage of the nation, and therefore as a topic in school and university curriculum. As for its social impact, it will contribute to the growing interest in those cultural dynamics which are affected and/or shaped by gender issues as well as promoting a deeper social awareness of the cultural role of women in Italian society.

As for its methodology, the actual collaboration among different scientific areas, such as DH, Computer Science, Cinema Studies, Photography, Visual Studies, Gender Studies, Literary Theory, will constitute an example of an integrated and multi-layered methodology, offering itself as a possible benchmark for potential future projects. As for its objectives and expected potential impacts, this work is in line with the research goals and targets defined by the National Research Program (PNR) 2021-2027 and by the "Cultural Heritage" specific intervention area of Horizon Europe. In particular, it will mainly impact the cluster *"Humanistic culture, creativity, social transformation, society of inclusion"* of the PNR, specifically the sub-category *"Digital preservation and conservation of cultural heritage"*. In fact, according to this sub-category research line, the priority should be given to the implementation of effective semantic modeling technologies, also through the construction of ontologies wherever necessary, which allow an effective aggregation of different information levels and types of data, in order to avoid redundancies or lack of data. The research should also include a shift from traditional databases to SW databases, in order to achieve interoperability between resources and set up the field for the transition to Big Data. All these research aspects will be considered within the development of the WOW Semantic Archive that will have an expected impact in supporting the digital transformation of the cultural sector, following the suggested focuses by the PNR related to *Digitization and Valorization of the Cultural Heritage*, but also strengthening the social inclusion, by reducing gender inequalities.

[5] https://www.europeanfilmgateway.eu/it

References

[1] E. Gardiner, R. G. Musto, The Digital Humanities: A Primer for Students and Scholars, Cambridge University Press, 2015.

[2] E. Hyvönen, Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193.

[3] F. Moretti, Distant reading, Verso Books, 2013.

[4] F. Ciotti, Modelli e metodi computazionali per la critica letteraria: lo stato dell'arte (2017).

[5] F. Ciotti, Distant reading in literary studies: a methodology in quest of theory, Testo e Senso (2021) 195–213.

[6] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, Scientific American 284 (2001) 34–43.

[7] C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, Linked Data on the Web (LDOW2008), in: Proceedings of the 17th international conference on World Wide Web, 2008, pp. 1265–1266.

[8] L. Pandolfo, L. Pulina, ARKIVO Dataset: A benchmark for ontology-based extraction tools., in: Proceedings of the 17th International Conference on Web Information Systems and Technologies, WEBIST 2021, October 26-28, 2021, SCITEPRESS, 2021, pp. 341–345.

[9] L. Pandolfo, S. Spanu, L. Pulina, E. Grosso, Understanding and modeling visitor behaviours for enhancing personalized cultural experiences, Int. J. Technol. Hum. Interact. 16 (2020) 24–38.

[10] N. Guarino, D. Oberle, S. Staab, What is an ontology?, in: Handbook on Ontologies, Springer, 2009, pp. 1–17.

[11] G. Adorni, M. Maratea, L. Pandolfo, L. Pulina, An ontology for historical research documents, in: Web Reasoning and Rule Systems - 9th International Conference, RR 2015, Berlin, Germany, August 4-5, 2015, Proceedings, volume 9209 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 11–18.

[12] G. Adorni, M. Maratea, L. Pandolfo, L. Pulina, An ontology-based archive for historical research, in: Proceedings of the 28th International Workshop on Description Logics, Athens,Greece, June 7-10, 2015, volume 1350 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015.

[13] L. Pandolfo, L. Pulina, M. Zieliński, Towards an ontology for describing archival resources, in: Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017, volume 2014 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 111–116.

[14] L. Pandolfo, L. Pulina, M. Zieliński, ARKIVO: an ontology for describing archival resources, in: P. Felli, M. Montali (Eds.), Proceedings of the 33rd Italian Conference on Computational Logic, Bolzano, Italy, September 20-22, 2018, volume 2214 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 112–116.

[15] L. Pandolfo, L. Pulina, M. Zieliński, Exploring semantic archival collections: The case of Piłsudski Institute of America, in: Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings, volume 988 of *Communications in Computer and Information Science*, Springer, 2019, pp. 107–121.

[16] L. Pandolfo, L. Pulina, Building the semantic layer of the Józef Piłsudski digital archive with an ontology-based approach, International Journal on Semantic Web and Information Systems (IJSWIS) 17 (2021) 1–21.

[17] M. Rizzarelli, L'attrice che scrive, la scrittrice che recita. per una mappa della 'diva-grafia', Vaghe stelle. Attrici del/nel cinema italiano. Arabeschi, edited by Lucia Cardone, Giovanna Maina, Stefania Rimini, and Chiara Tognolotti 10 (2017) 366–371.

[18] G. Simi, L'occhio che palpita. monica vitti e gli scritti sull'arte, Cinergie–Il Cinema e le altre Arti (2021) 153–166.

[19] M. Rizzarelli, Il doppio talento dell'attrice che scrive. per una mappa delle "divagrafie", Cahiers d'études italiennes (2021).

[20] C. Tognolotti, Una diva fragrante. l'immagine divistica di sophia loren nei libri di ricette (2019).

[21] L. Pandolfo, L. Pulina, G. Adorni, A framework for automatic population of ontology-based digital libraries, in: AI*IA 2016: Advances in Artificial Intelligence - XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 - December 1, 2016, Proceedings, volume 10037 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 406–417.

[22] ADnOTO: a self-adaptive system for automatic ontology-based annotation of unstructured documents, in: Advances in Artificial Intelligence: From Theory to Practice - 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part I, volume 10350 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 495–501.