Information Organization and Access in Digital Humanities

TaDiRAH Revised, Formalized and FAIR

Luise Borek

Technical University of Darmstadt, Germany
luise.borek@tu-darmstadt.de

Canan Hastik

DIPF | Leibniz Institute for Research and Information in Education, Frankfurt/M., Germany
hastik@dipf.de

Vera Khramova

Darmstadt University of Applied Sciences, Germany
vera.khramova@stud.h-da.de

Klaus Illmayer

Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), Vienna, Austria
klaus.illmayer@oeaw.ac.at

Jonathan D. Geiger

Academy of Sciences and Literature, Mainz, Germany
jonathan.geiger@adwmainz.de

Abstract

Classifying and categorizing the activities that comprise the digital humanities (DH) has been a longstanding area of interest for many practitioners in this field, fueled by ongoing attempts to define the field both within the academic and public sphere. Several European initiatives are currently shaping advanced research infrastructures that would benefit from an implementation of a suiting taxonomy. Therefore, new humanities and information science collaborations have been formed to provide a service that meets their needs. This working paper presents the transformation of the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) in order to make it machine-readable and become a formalized taxonomy. This includes the methodology and realization containing a complete revision of the original version, decisions in modelling, the implementation as well as organization of ongoing and future tasks. TaDiRAH addresses a wide range of humanities disciplines and integrates application areas from philologies as well as epigraphy, and musicology to name just a few. For this reason, the decision in favor of SKOS was made purely pragmatically in terms of technology, concept and domains. New language versions can now be easily integrated and low-threshold term extensions can be carried out via Wikidata. The new TaDiRAH not only represents a knowledge organization system (KOS) which has recently been released as version 2.0. According to the FAIR principles this new version improves the Findability, Accessibility, Interoperability, and Reuse of research data and digital assets in the digital humanities.

Keywords: taxonomy; digital humanities; linked open data; knowledge organization

1 Motivation

The knowledge organization within the digital humanities is located at the intersection of scientific practice and state of the art. The investigation of the genesis of this domain shows that knowledge modelling has become established as a standard procedure (Unsworth, 2000b; McCarty, 2003; Schnapp, 2011) not only in text-oriented sciences. This practice integrates both subject-specific and information science methods that are generally known as 'scholarly primitives' within the digital humanities domain (Unsworth, 2000a). While the objects of research and their materiality vary greatly in the participating disciplines, 'methodological commons' may be identified, under which generic and more specific research activities may be subsumed.

For a better understanding of the general digital scientific process and for the shaping of knowledge, new knowledge models are being continuously developed and expanded. The Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) is such a model and was designed in close collaboration with the community to organize and categorize digital humanities content and generate academic credibility as well as greater visibility to the field (Borek et al., 2016). At the same time, it has become an instrument to reflect upon digital humanities as a discipline as its broader implementation provides insights on which activities are being used in which contexts. In this sense, the taxonomy is never complete and continuously depends on the community to help shape and adapt it according to their needs.

Since 2015 this vocabulary is used in a variety of contexts to categorize bibliographic data (DARIAH Zotero bibliography "Doing Digital Humanities"[1]), research projects (e.g., in the registries DHCommons[2] and AGATE European Science Academies Gateway for the Humanities and Social Sciences[3]), study programmes (Digital Humanities Course Registry[4]), conference abstracts, and research tools (TAPoR Text Analysis Portal for Research[5], CLARIAH-DE[6], and the SSH Open Marketplace[7]).

From the beginning, TaDiRAH has been a community-driven initiative that pursues a practice-oriented approach, in which it benefits from further contributions, collaborations, development and reuse. The interdisciplinary community of digital humanities scholars have not only helped shape the relevant terminology and the associated definitions, they have also provided translations on concept level to several languages such as German, French, Spanish, Portuguese, Serbian, and quite up-to-date Italian.

While TaDiRAH has been in wide use over the years, its application had not gone beyond keywording as the non-funding project did not have the necessary resources to provide a formalized, persistent and machine-readable version. With several European initiatives currently shaping advanced research infrastructures that could benefit from an implementation of the taxonomy, new humanities and information science collaborations have been formed to bring TaDiRAH to its next level.

In this paper, we describe the steps that were undertaken in order to transfer TaDiRAH to a machine readable taxonomy as a knowledge organization system (KOS) currently released as version 2.0 within the Vocabs service at the Austrian Academy of Sciences.[8] This working paper covers the methodology and realization that includes a complete revision of the original version, decisions in modelling, the implementation as well as the organization of ongoing and future tasks.

[1] Doing Digital Humanities – A DARIAH Bibliography, https://www.zotero.org/groups/113737/doing_digital_humanities_-_a_dariah_bibliography (Nov. 18, 2020).

[2] now within https://dhcenternet.org/ (Nov. 18, 2020)

[3] AGATE. A European Gateway for the Academies of Sciences and Humanities, https://agate.academy/de.html (Nov. 18, 2020)

[4] https://dhcr.clarin-dariah.eu/ (Nov. 18, 2020)

[5] TAPoR 3. Discover research tools for studying texts, http://tapor.ca (Nov. 18, 2020)

[6] https://www.clariah.de/ (Nov. 18, 2020)

[7] SSH Open Marketplace, https://www.sshopencloud.eu/ssh-open-marketplace (Nov. 18, 2020)

[8] TaDiRAH: Taxonomy of Digital Research Activities for the Humanities, Version 2.0.0, https://vocabs.dariah.eu/tadirah/ (Nov. 18, 2020)

2 Methodology

In information science, a term represents an abstract concept for a set of objects and items with common properties. This term is linguistically represented by a name. These designations can be natural language words but also a notation of a classification or order of knowledge. Various methods can be used to introduce these concepts into taxonomy. An empirical procedure for the concept analysis would be the semi-automatic calculation of word similarities in texts by means of cluster analysis. Therefore, a representative reference corpus must already be available. The method of formal concept analysis is used to determine connections and hierarchical relationships between concepts. A tree-like order structure is designed using superordinate and subordinate conceptual relations, with which common properties are combined into concepts. Following the hermeneutic approach, terms are subject to change and this change must always be viewed in its context. In pragmatics, the meaning, purpose and aim of terms are in the foreground while language always influences conceptualization (Hjørland, 2009).

The first TaDiRAH version 0.5.1[9] with its research activities, techniques and objects was created through a case study in pragmatic classification based on a hermeneutical and iterative approach intensively discussed with the community (Borek et al., 2016). The resulting concept order comprises three independent concept trees, each with one superordinate and various subordinate levels. The term definitions contain information about synonyms and their semantic context but also minimal deviations from adjacent terms.

In order to eliminate the vagueness and to optimize the usefulness of the existing vocabulary for the growing knowledge domain of digital humanities, the need arose to revise and harmonize the structure and semantics of the model, as well as the concept terms and definitions, the so called scope notes, with respect to research activities and research techniques. This also includes expanding the model on the basis of community usage. As a result, a focus has been set on the research activities and the relation of techniques associated to them. To achieve this, existing implementations such as TAPoR, SSH und CLARIAH-DE were evaluated and aggregated in the model. At the same time, this led to the decision to exclude related research objects from

[9] TaDiRAH: Taxonomy of Digital Research Activities for the Humanities, http://tadirah.dariah.eu/vocab/index.php (Nov. 18, 2020)

the model as no reliable systematic data has been available for modelling. However, research objects may be associated through suiting vocabularies in the future (within TaDiRAH or from additional sources).

In addition, criteria for the descriptions have been established to develop consistent definitions of terms with a focus on their usefulness for the domain. Scope notes should in future not be too narrow nor too broad, and should not contain any or explicit demarcations to other terms. Moreover, scope notes should not include contradictions or negative definitions. In principle, multiple definitions within a scope note, redundancies with other definitions and unnecessary references should be avoided. Finally, TaDiRAH terms should be consistently written in small letters and gerundiva to represent the performativity of the activities. This is to be largely continued in the subclasses. In addition to the scope notes that are assigned to the TaDiRAH core model on top and sub-concept level, so-called 'aggregated concepts' – mostly represented by research techniques – are extended by external Wikidata definitions.

Only the linking of the totality of all concepts with each other allows for a knowledge order to emerge from a concept definition. These quasi hardwired terms and their semantic relations provide an overview of the hierarchical, associative relationship between terms and their corresponding expressiveness of content in the resulting knowledge model and enable an improved search and retrieval. While it is required for a vocabulary to be consistent and coherent, it does not necessarily have to be complete, in order to perform search and retrieval. A taxonomy must always be open for extensions without the need to revise existing definitions (Gruber, 1993). This modelling process integrates analysis and organization activities while the next process step indicates a strong relation to knowledge engineering methods by formalizing and implementing the TaDiRAH conceptualization in standardized Simple Knowledge Organization System (SKOS)[10]. As a specification and standard according to the FAIR principles[11] SKOS and the digital infrastructure where it is published makes TaDiRAH findable, accessible, interoperable and reusable as a knowledge organization system.

[10] SKOS Simple Knowledge Organization System, https://www.w3.org/2004/02/skos/ (Nov. 18, 2020)

[11] FAIR Principles, https://www.go-fair.org/fair-principles/ (Nov. 18, 2020)

3 Modelling and implementation

Besides the redesign of TaDiRAH and the transformation of a pior nomenclature into a standardized taxonomy, its SKOSification was also implemented in several iterations. The goal was to make the simple keyword list of different research activities, research techniques and research objects, used in the digital humanities until then, interoperable available in a machine-readable form and to further improve the search for relevant information by semantically linking these data. This structured representation of the data and the relationships between the data is an important prerequisite to demonstrate the applicability and range of the developed model, but also for its use and reuse.

The machine-readable form will be achieved with the help of SKOS. As a formal language based on the Resource Description Framework (RDF)[12], SKOS with its clear structure and low level of detail allows easy to carry out adjustments and extensions of the model. SKOS was developed specifically for the representation of data of controlled vocabularies and taxonomies. It can be integrated into the Semantic Web and the modelled data can be stored as Linked Open Data (LOD) and further be linked to external resources. In this way, not only the availability and visibility of knowledge can be improved, but also interoperability with related disciplines can be achieved. For the creation of the SKOSified TaDiRAH model, Protegé[13] was used, an ontology editor that facilitates the modelling process and also provides a visual representation of the model.

3.1 SKOSification: Concept hierarchy and semantification

Currently the revised TaDiRAH metadata profile includes 168 terms. The task included the creation of a model that would be broad enough to cover the most important research activities in the digital humanities but would be specific enough to help structure various digital humanities tools, projects, websites and bibliographies.

[12] W3C Recommendation 25 February 2014, https://www.w3.org/TR/rdf11-concepts (Nov. 18, 2020)

[13] Protégé v.5.5.0, a free, open-source ontology editor and framework for building intelligent systems, from https://protege.stanford.edu/ (Nov. 18, 2020)

All terms in the scheme are represented as concepts by **skos:Concept** which is the central structural element in SKOS. Each concept is unique and identified by a Uniform Resource Identifier (URI), thus ensuring that each resource is uniquely identifiable on the web. All individuals must belong to a specific concept scheme, which is expressed through the class **skos:ConceptScheme**. Moreover, they are represented with a scope note or an external Wikidata definition, some broader and narrower terms, and also assigned to multilingual labels, currently a total of seven languages. Since the translations are based on the 'old' TaDiRAH version, they are neither complete nor consistent, and should be revised and adapted to the new model version 2 as soon as possible. All properties are expressed in the form of relations to represent the individual concept as well as its interrelation with other concepts. An example of the SKOS structure of TaDiRAH version 2 is shown in the next figure (Fig. 1).

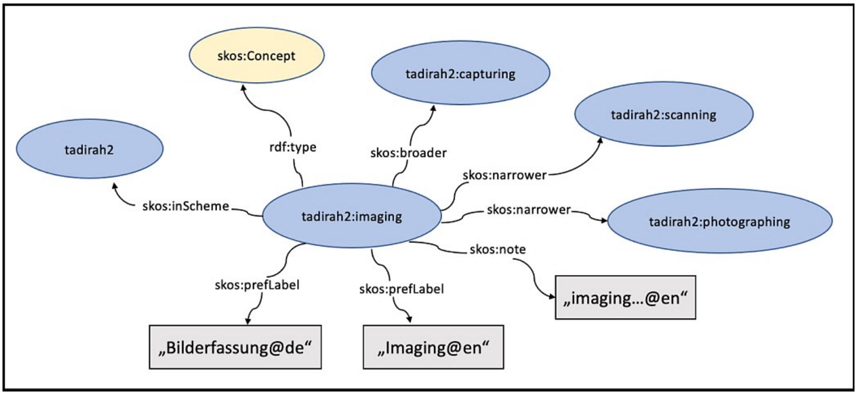
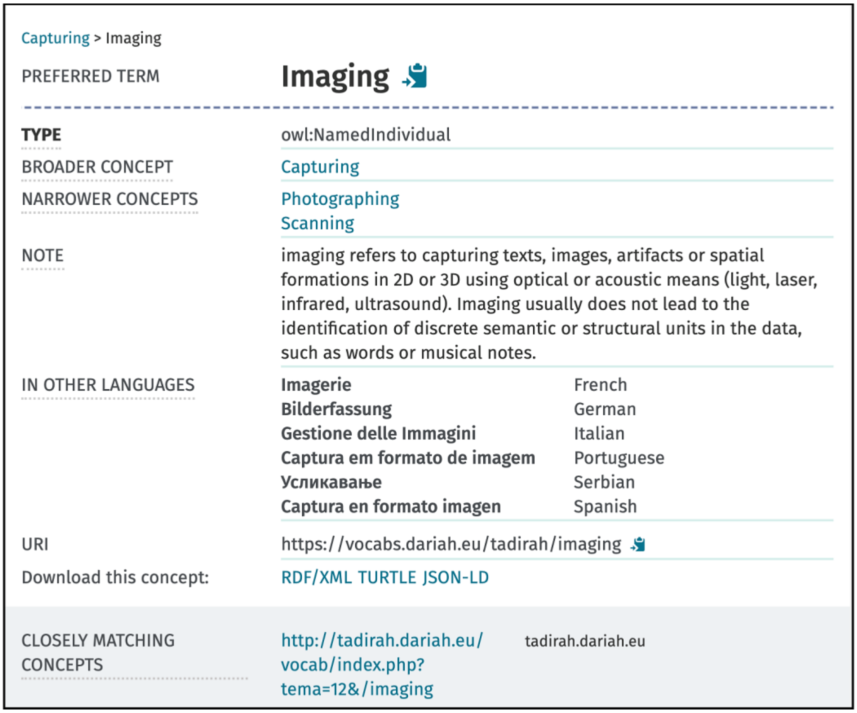


Fig. 1 SKOS structure of the concept *Imaging*

Hierarchical conceptual order between the scheme and the top concepts is achieved using the properties **skos:hasTopConcept** and **skos:topConceptOf**. From a total of 168 terms only seven are top concepts, namely *Analyzing*, *Capturing*, *Creating*, *Disseminating*, *Enriching*, *Interpreting* and *Storing*. The affiliation of all concepts to the schema is made recognizable with the property **skos:inScheme**. To represent the hierarchical relationships between the various concepts the properties **skos:broader** and **skos:narrowwer** are used. The properties **skos:prefLabe**l and **skos:altLabel** are required to indicate preferred and alternative labels of the resources. To express the multilingual labels the property **skos:prefLabel** is used. Information about the meaning of a resource is held by the properties **skos:scopeNote** or **skos:definition**. These two properties serve to distinguish the self-written scope notes from external Wikidata definitions, which are identified with the property **skos:definition**. In addition, the property **skos:closeMatch** is needed to create the mapping between the first and the new TaDiRAH version, but also through the terms taken from Tapor, SSH and CLARIAH-DE (e.g., *Optical Character Recognition*, *Searching*, *Archiving*) which are successfully aggregated in the model.

The new features outlined here simultaneously mean formalization and a semantic enrichment that improve the taxonomy and facilitate its application. The SKOSification resolved former ambiguity and vagueness, transformed implicit information into explicit representation, and led to more flexibility.

Prior only hierarchically organized concepts have been transformed into a formal classification, thus enabling semantic interpretation of concepts, their relations and represented entities.



Capturing > Imaging

PREFERRED TERM	**Imaging**	
TYPE	owl:NamedIndividual	
BROADER CONCEPT	Capturing	
NARROWER CONCEPTS	Photographing Scanning	
NOTE	imaging refers to capturing texts, images, artifacts or spatial formations in 2D or 3D using optical or acoustic means (light, laser, infrared, ultrasound). Imaging usually does not lead to the identification of discrete semantic or structural units in the data, such as words or musical notes.	
IN OTHER LANGUAGES	**Imagerie**	French
	Bilderfassung	German
	Gestione delle Immagini	Italian
	Captura em formato de imagem	Portuguese
	Усликавање	Serbian
	Captura en formato imagen	Spanish
URI	https://vocabs.dariah.eu/tadirah/imaging	
Download this concept:	RDF/XML TURTLE JSON-LD	
CLOSELY MATCHING CONCEPTS	http://tadirah.dariah.eu/vocab/index.php?tema=12&/imaging	tadirah.dariah.eu

Fig. 2 Concept – *Imaging*, from https://vocabs.dariah.eu/tadirah/imaging (Nov. 18, 2020)

3.2 Implementation and publication with Vocabs

The TaDiRAH vocabulary is published on the vocabulary server of DARIAH-EU.[14] This service is run by the Austrian Centre for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences, one of the partners in the CLARIAH-AT consortium.[15] For showing and browsing the vocabularies the open source software Skosmos[16] is used. This software works natively on a triple store database where it expects a SKOS compatible vocabulary to then show in the browser. The triple store that is used is an Apache Jena Fuseki setup.[17] Usually, the workflow for inserting and updating a vocabulary expects an RDF import file, e.g., in a Turtle format. Such a file is manually imported into Apache Jena Fuseki in a dedicated namespace. This namespace is also declared in the configuration file of Skosmos and after a successful import the changes are immediately taken over by Skosmos. There is also an interface for SPARQL if there is a need for complex queries.[18] This service is used by a growing number of vocabularies, focusing on such vocabularies that are used in the DARIAH community (mainly research disciplines out of the humanities and arts). This governance model guarantees sustainability and maintainability of the service and the data of the vocabularies.

Skosmos allows easy browsing in a vocabulary, it enables a search option and it does in general present the concepts of a vocabulary in a structured, human-readable way. It enables an easy way to interact with a SKOS based vocabulary and therefore supports dissemination and usage of the TaDiRAH vocabulary. The web server is set up in a way that the concept URIs dissolves as expected. Relations between concepts of the new and the first version of TaDiRAH allow easy mapping and give a good overview on the differences between the versions. Changes at the new version are communicated by a

[14] Vocabulary services (Vocabs), https://vocabs.dariah.eu and general information on DARIAH-EU, https://www.dariah.eu (Nov. 18, 2020)

[15] Information on CLARIAH-AT, https://clariah.at (Nov. 18, 2020)

[16] Download Skosmos, http://skosmos.org (Nov. 18, 2020)

[17] Apache Jena Fuseki, https://jena.apache.org/documentation/fuseki2/index.html (Nov. 18, 2020)

[18] For more information on the SPARQL interface, https://vocabs.dariah.eu/en/about#sparql (Nov. 18, 2020).

versioning system.[19] The approach is to be as open and as transparent as possible. Therefore the RDF export of the vocabulary can be downloaded in a turtle format and due to the licensing of TaDiRAH vocabulary under the Creative Commons public domain license (CC0 1.0) reuse is highly appreciated.[20] All in all, the publication settings take care that the TaDiRAH vocabulary conforms to the FAIR data principles, such as the use of URIs, to uniquely identify the concepts, the providing of machine readable form of the model, using W3C Standard SKOS and the usage of CC0 1.0 licence.

4 Future ongoing tasks

After the revision and publication of the new TaDiRAH version, a workflow for further maintenance, development, supervision, community engagement, monitoring, and quality management has been designed and implemented.

A TaDiRAH board will be responsible for process management in the future.[21] The board consists of the original core team, new developers and other contributors. The future tasks include the maintenance of the published model, for example by correcting errors and editing GitHub-issues[22], and the further development of the model according to the needs and commitment of the community. In addition, the board will coordinate other translations, such as the Norwegian language, which is currently being developed, and will ensure technical and professional mentoring for the use and reuse of the model. Another important task of the board is the continuous documentation and dissemination of current information via the website[23] providing visualization of the model, but also in publications and via social media[24]. Project

[19] Using the relation **owl:versionInfo**, the version numbers are based on Semantic Versioning 2.0.0, see: https://semver.org/ – additionally the relation **dct:modified** informs about the date of the last change.

[20] CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, https://creativecommons.org/publicdomain/zero/1.0/ (Nov. 18, 2020)

[21] See TaDiRAH website for further activities: https://tadirah.info/ (Jan. 15, 2021).

[22] https://github.com/dhtaxonomy/TaDiRAH (Jan. 15, 2021)

[23] TaDiRAH as Linked Open Data, https://tadirah.info/ (Nov. 18, 2020)

[24] https://twitter.com/tadirah_dh (Nov. 18, 2020)

and publication lists will also be constantly maintained. Finally, the quality of the new TaDiRAH model depends on the commitment and involvement of the community, which is invited to participate in the revision of the multilingual terminology and scope notes, to submit new GitHub issues, and to continue the list of applications, users, and publications.

The community is also invited to further develop the model in terms of content and structure. For this purpose, an approach has been created which, by interlinking with Wikidata, enables a workflow in which missing definitions are iteratively supplemented, corrected and ideally transformed into high-quality scope notes. Current aggregated concepts link to Wikidata items, which partly do not yet contain any or only rudimentary definitions. At this point, the model can be further extended.

Another task for the future of TaDiRAH is to deepen the relationship between TaDiRAH and Wikidata. That means on the one hand to add definitions of terms or concepts provided by the taxonomy but not by Wikidata or to improve existing but deficient definitions on Wikidata. This will improve Wikidata as a provider of linked open data and expand its scope towards terms highly relevant within the domain of the digital humanities. On the other hand, that means to make good use of precise definitions given in Wikidata for TaDiRAH especially for its scope notes. This two-folded contrastive method will not only improve collections of controlled terms but also inspire further development of TaDiRAH as ideas for new concepts, terms, relations, or their reshaping, may be provoked. This will be a good step to link TaDiRAH and Wikidata data-wise and methodically and might serve as a best practice example for similar approaches.

References

Borek, L., Dombrowski, Q., Perkins, J., & Schöch, C. (2016). TaDiRAH – A Case Study in Pragmatic Classification. *Digital Humanities Quarterly* (*DHQ), 10*(1). http://digitalhumanities.org/dhq/ vol/10/1/000235/000235.html

Gruber, Thomas R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition, 5*(2), 199–220.

Hjørland, Birger (2009). Concept theory. *Journal of the American Society for Information Science and Technology, 60*(8), 1519–1536.

McCarty, W. (2003). Humanities Computing. In Drake, M. (Ed.), *Encyclopedia of Library and Information Science* (2nd ed., pp. 1224–1235). New York: Marcel Dekker.

Schnapp, Jeffrey T. (2011). Emerging Disciplines ⇒ Knowledge Design. http://jeffreyschnapp.com/2011/08/26/finding-new-field-descriptions/

Suominen, O., Ylikotila, H., Pessala, S., Lappalainen, M., Frosterus, M., Tuominen, J. … Retterath, A (2015). Publishing SKOS vocabularies with Skosmos. http://skosmos.org/ publishing-skos-vocabularies-with-skosmos.pdf

Unsworth, J. (2000a). Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html

Unsworth, J. (2000b). What is Humanities Computing and What is Not? http://www.people.virginia.edu/~jmu2m/mith.00.html

Zaytseva, K., & Ďurčo, M. (2020). Controlled Vocabularies and SKOS. Version 1.0.0. DARIAH-Campus [Training module]. https://campus.dariah.eu/resource/controlled-vocabularies-and-skos

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 321–332. DOI: doi.org/10.5283/epub.44951.