Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

Schriften des
Instituts für Dokumentologie und Editorik

Band 15

Schriften des Instituts für Dokumentologie und Editorik — Band 15

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

edited by

Elena Spadini, Francesca Tomasi, Georg Vogeler

2021

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 5. Dezember 2021.

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7543-4369-2
Einbandgestaltung: Stefan Dumont nach Vorarbeiten von Johanna Puhl und Katharina Weber
Satz: LuaTEX, Bernhard Assmann

Contents

Infrastructures and Technologies

Formal Models

Projects and Editions

Appendices

Preface

This volume is based on the selected papers presented at the *Workshop on Scholarly Digital Editions, Graph Data-Models and Semantic Web Technologies*, held at the University of Lausanne in June 2019. The *Workshop* was organized by Elena Spadini (University of Lausanne) and Francesca Tomasi (University of Bologna), and sponsored by the Swiss National Science Foundation through a Scientific Exchange grant, and by the Centre de recherche sur les lettres romandes of the University of Lausanne. The *Workshop* comprised two full days of vibrant discussions among the invited speakers, the authors of the selected papers, and other participants.[1] The acceptance rate following the open call for papers was around 60%. All authors – both selected and invited speakers – were asked to provide a short paper two months before the *Workshop*. The authors were then paired up, and each pair exchanged papers. Paired authors prepared questions for one another, which were to be addressed during the talks at the *Workshop*; in this way, conversations started well before the *Workshop* itself. After the *Workshop*, the papers underwent a second round of peer-review before inclusion in this volume. This time, the relevance of the papers was not under discussion, but reviewers were asked to appraise specific aspects of each contribution, such as its originality or level of innovation, its methodological accuracy and knowledge of the literature, as well as more formal parameters such as completeness, clarity, and coherence. The bibliography of all of the papers is collected in the public Zotero group library *GraphSDE2019*[2], which has been used to generate the reference list for each contribution in this volume.

The invited speakers came from a wide range of backgrounds (academic, commercial, and research institutions) and represented the different actors involved in the remediation of our cultural heritage in the form of graphs and/or in a semantic web environment. Georg Vogeler (University of Graz) and Ronald Haentjens Dekker (Royal Dutch Academy of Sciences, Humanities Cluster) brought the Digital Humanities research perspective; the work of Hans Cools and Roberta Laura Padlina (University of Basel, National Infrastructure for Editions), as well as of Tobias Schweizer and Sepideh Alassi (University of Basel, Digital Humanities Lab), focused on infrastructural challenges and the development of conceptual and software frameworks to support researchers' needs; Michele Pasin's contribution (Digital Science, Springer Nature) was informed by his experiences in both academic research, and in commercial technology companies that provide services for the scientific community.

[1] The archived website of the event, including links to the slides of the presentations, is available at <https://zenodo.org/record/3402173>

[2] Available at <https://www.zotero.org/groups/2339860/graphsde2019/library>.

The *Workshop* featured not only the papers of the selected authors and of the invited speakers, but also moments of discussion between interested participants. In addition to the common Q&A time, during the second day one entire session was allocated to working groups delving into topics that had emerged during the *Workshop*. Four working groups were created, with four to seven participants each, and each group presented a short report at the end of the session. Four themes were discussed: enhancing TEI from documents to data; ontologies for the Humanities; tools and infrastructures; and textual criticism. All of these themes are represented in this volume.

The *Workshop* would not have been of such high quality without the support of the members of its scientific committee: Gioele Barabucci, Fabio Ciotti, Claire Clivaz, Marion Rivoal, Greta Franzini, Simon Gabay, Daniel Maggetti, Frederike Neuber, Elena Pierazzo, Davide Picca, Michael Piotrowski, Matteo Romanello, Maïeul Rouquette, Elena Spadini, Francesca Tomasi, Aris Xanthos – and, of course, the support of all the colleagues and administrative staff in Lausanne, who helped the *Workshop* to become a reality.

The final versions of these papers underwent a single-blind peer review process. We want to thank the reviewers: Helena Bermudez Sabel, Arianna Ciula, Marilena Daquino, Richard Hadden, Daniel Jeller, Tiziana Mancinelli, Davide Picca, Michael Piotrowski, Patrick Sahle, Raffaele Viglianti, Joris van Zundert, and others who preferred not to be named personally. Your input enhanced the quality of the volume significantly!

It is sad news that Hans Cools passed away during the production of the volume. We are proud to document a recent state of his work and will miss him and his ability to implement the vision of a digital scholarly edition based on graph data-models and semantic web technologies.

The production of the volume would not have been possible without the thorough copy-editing and proof reading by Lucy Emmerson and the support of the IDE team, in particular Bernhard Assmann, the TeX-master himself. This volume is sponsored by the University of Bologna and by the University of Lausanne.

Bologna, Lausanne, Graz, July 2021

Francesca Tomasi, Elena Spadini, Georg Vogeler

Introduction

Elena Spadini, Francesca Tomasi

This volume explores possible interactions between digital texts, the graph data-model, scholarly editions, and the semantic web. Combinations of these objects and concepts, first explored in the past decades, remain experimental to date, and represent one possible area of development for the field of digital scholarly editing.

The conceptual and technological stack adopted in most digital editions today[1], as well as in the tools, software, and platforms used to produce them, is based on the centrality of the document, on hyperlinks within closed databases, and on hierarchical text encoding. This *status quo* is challenged by the need to represent multiple perspectives on the same textual work, and by the diffusion of knowledge graphs to organize the relationships between data in a semantically explicit and computationally readable way.

Our experience in creating what we call today a Semantic Scholarly Digital Edition (SSDE) started with the project *Vespasiano da Bisticci, Lettere* (Tomasi 2013), where, in the first release, we proposed a model for managing a documentary approach to the edition through semantic web technologies (Tomasi 2012). Working on the notion of 'knowledge site', by the time the edition reached the third version, it proposed a complete Linked Open Data publication of the collection of letters (Tomasi 2020), by also focusing on the topic of reproducibility of an SSDE[2], as a way to enable dialogue within the scholarly community.

Scholarly edition of historical documents have often been pioneering in exploring the possibilities offered by the graph data-model, and by semantic web technologies (Meroño-Peñuela et al. 2014). In a process of "extension of indices" (Vogeler 2020, 44), they make use of available conceptual models and ontologies (e.g., CIDOC-CRM and FOAF) for the modelling of named entities like events, persons, and places, particularly relevant in this kind of editions.

In scholarly editing, the identification and linking of semantic entities often needs to be coupled with the representation of other aspects of a textual work, such as intertextuality, style, textual variation, as well as the expression of linguistic, palaeographic and codicological features. Scholarly editors and digital humanists are not yet equipped to represent these phenomena in the growing world of knowledge graphs, because of the relative lack of dedicated formal models, standards, and pieces of

[1] Good catalogues of digital editions are available in Franzini et al. (2016-) and Sahle (2008-).

[2] See Tomasi 2020, section *Documentation* available at <https://dharc-org.github.io/vespasiano-da-bisticci-letters-de/documentation/index.html>.

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 1–6.

infrastructure[3]. The reflections and contributions in these proceedings aim to address these needs. Towards this same end, for instance, we proposed a model for the relationships between the different entities concerned by genetic criticism (documents, publications, dossiers), in the context of the scholarly edition of the complete works of Gustave Roud (Christen and Spadini 2019).

Some challenges are indeed common to all kinds of scholarly editions – literary, historical or others. Among them is the need to represent the editorial assertions, or the interpretation, as explicitly as possible, and the requirement for referencing each element (portion of the text, interpretation act, or external entity) in a unique and permanent manner. Provenance-aware models and standards for machine consumption of textual resources are fundamental to addressing these demands, and both topics are well represented in this volume.

The experiment we carried out on Bufalini's notebooks (Daquino et al. 2019a) pursued these same issues, and gave us the opportunity to propose a complete workflow for SSDEs.[4] Starting from the text transcription (a non-neutral action, which required the modelling of both the digital edition features and the XML/TEI markup), we then moved onto the data modelling phase, bearing in mind the need to reuse existing ontologies, and the importance of assuming a provenance-aware approach. We adopted the same approach for the RDF transformation, which required us to use named graphs and the nanopublication model (Groth et al. 2010). Finally, we developed a web application, which was useful not only to serve and exploit RDF data, but also to abide by the principle of using a data-driven visualization to enhance user-experience.

There is still a lot of work to do in order to more-fully exploit the possibilities of this vast domain, and this collection of papers aims to face these demanding challenge. With this volume, we hope to propose some interesting suggestions for doing research in the field of graph data-models and semantic web technologies in scholarly digital editing.

The topics addressed in this proceedings explore this subject from different perspectives: of the infrastructures, of the formal models, and of real-life project implementations. The first of these perspectives, that is, the need for further infrastructure

[3] The graph data-model has been successful in the representation of textual variants (Schmidt and Colomb 2009), notably thanks to the Gothenburg model for automatic collation, but the related MVD (multi-version document) format has not been widely adopted. For what concerns formal models, for citations and intertextuality see the Citation Typing Ontology (CiTO) (Peroni and Shotton 2012) and the Intertextual Relationships Ontology for literary studies (INTRO) (Oberreither 2020); for linguistics, cf. for example Chiarcos et al. 2018; Tittel et al. 2018; Franzini 2019; for ancient and medieval documents, see Sharing Ancient Wisdoms Ontology (SAWS) (2012), Linked Ancient World Data Ontology (LAWD) (Cayless 2014) and CRMtex (Murano and Felicetti 2020).

[4] See Daquino et al. 2019b, section *Introduction* available at <http://projects.dharc.unibo.it/bufalini-notebook/introduction>.

to cover existing gaps, and advancements that are being made in this domain, are explored in this volume by the contributions of Boot and Koolen, Cayless and Romanello, Neill and Schmidt, Prosser and Schloen, and Vogeler. They touch upon long-standing issues as diversified as APIs for texts, friendly GUIs, the integration of TEI and RDF, data granularity, and technologies affordance. Boot and Koolen consider the question of how to derive RDF triples from a TEI encoded text and where to store them; their edition server serves the triples to any annotation tool, in order for users to analyse or enrich an edition. The data are modelled according to an ontology for the editorial domain. The paper by Cayless and Romanello addresses the gap in the available digital infrastructure of services for handling the resolution of URIs for texts and their parts. Some real-life uses cases for such services are introduced, and a proposal of their three main components – a registry service, an identifier resolution service and a document metadata scheme – is presented and documented with examples. Neill and Schmidt introduce the SPEEDy editor, its data model for annotations, and its user interface. SPEEDy is a standoff editor, through which any range of text, including overlapping ones, can be enriched with one or more annotations. Prosser and Schloen describe the OCHRE system, which has been successfully used for around a decade to manage text corpus projects. OCHRE, built as a graph database, is flexible and customizable, thanks to a semi-structured data model in which atomized data is made available for different kinds of arrangements and links. The paper by Vogeler focuses on the technological affordances of the XML hierarchical tree model, the RDF graph representation, and stand-off annotations, exploring the potential of virtuously combining the three.

The second perspective, concerning the question of formal models, is central to the contribution of Giovannetti, as well as of Cools and Padlina. These articles share with others a scalable approach to ontology development: overarching conceptual models are defined or reused, and guide the creation of domain- or project-specific models. Among the former, CIDOC-CRM and FRBR form a reference point for representing the semantics of events, cultural objects, and their relationships. This does not mean, however, that they cover the requirements of any scholarly edition: their use is still experimental in this domain, as is shown by the CRMtex (Murano & Felicetti 2020), an extension of the CIDOC-CRM that supports the study of ancient documents, the first version of which was only released in June 2020. This volume should help to expose some of the advantages and limitations of the use of these higher-level ontologies. Cools and Padlina provide a description of the infrastructure using semantic web technologies created by the Swiss National Infrastructure for Editions (NIE-INE, University of Basel) in order to manage eleven digital edition projects. The focus of the paper is on data modelling, to which a scalable approach is applied in order to integrate generic and project-specific ontologies. The contribution by Giovannetti proposes an ontology for capturing the critical apparatus as a knowledge graph, which enables, by the use of formal semantics, the decoupling of text and interpretation, and the integration of scholarly editions representing textual variance in the linked open data cloud.

The articles by Burrows et al., by Münnich and Ahrend, and by Sippl, Burghardt and Wolff take the third perspective, each reporting on very different real-life projects, from manuscript description to music and medieval charters. All of the scholars involved in these projects (all of which make use of semantic web technologies) saw the benefit of taking a data-centric approach to scholarly editing and manuscript studies in order to enhance the value of the relationships between different kinds of data, as well as among multiple datasets. The paper by Burrows et al. reports on the results of the transformation of the TEI encoded manuscripts catalogues of the Bodleian Library into Linked Open Data. Special attention is devoted to the provenance data, which are only semi-structured in TEI. The paper by Münnich and Ahrend offers a compelling state of the art of the formalization of music as a graph, before providing an overview of the modelling approach chosen in the *Anton Webern Gesamtausgabe*, covering semantic relationships between different areas of the edition, and interoperability with other datasets. The paper by Sippl, Burghardt and Wolff focuses on a digital edition of medieval charters, which is based on an extremely heterogeneous dataset. NLP techniques are used for entities extraction. The graph database chosen is Neo4j, on top of which an exploratory web application is developed. The data model, which makes substantial use of CIDOC-CRM, is illustrated in detail.

To conclude, we believe this volume will be a valuable resource for anyone embarking on the adventure of joining scholarly editing with graph data models both inside and outside the emerging Linked Open Data ecosystem.

Bibliography

Cayless, Hugh, *Linking Ancient World Data Ontology (LAWD)*, 2014 <http://lawd.info/>

Chiarcos, Christian, Émilie Pagé-Perron, Ilya Khait, Niko Schenk, and Lucas Reckling, 'Towards a Linked Open Data Edition of Sumerian Corpora', in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan: European Language Resources Association (ELRA), 2018) <https://www.aclweb.org/anthology/L18-1387>

Christen, Alessio, and Elena Spadini, 'Modeling Genetic Networks. Gustave Roud's Œuvre, from Diary to Poetry Collections', *Umanistica Digitale*, 7 (2019) <https://doi.org/10.6092/issn.2532-8816/9063>

Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi, 'Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell'edizione semantica del quaderno di appunti di Paolo Bufalini', *Umanistica Digitale*, 3.7 (2019) <https://doi.org/10.6092/issn.2532-8816/9091>

Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi, eds., *Paolo Bufalini's Notebook, 1981-1991. The Digital Edition.* (Università di Bologna: /DH.arc, 2019) <http://projects.dharc.unibo.it/bufalini-notebook>

Franzini, Greta, 'Towards Connecting Scholarly Editions to Corpora in the LiLa (Linking Latin) Knowledge Base of Linguistic Resources', in *Lesen und lesen lassen. Edition und Korpus*, Vortragsreihe des IZED im Wintersemester 2019/20 (Wuppertal (Germany): University of Wuppertal, 2019) <https://doi.org/10.5281/zenodo.3613371>

Franzini, Greta, Peter Andorfer, and Ksenia Zaytseva, *Catalogue of Digital Editions: The Web Application*, 2016 <https://doi.org/10.5281/zenodo.1250796>

Groth, Paul, Andrew Gibson, and Jan Velterop, 'The Anatomy of a Nanopublication', *Information Services & Use*, 30.1–2 (2010), 51–56 <https://doi.org/10.3233/ISU-2010-0613>

Meroño-Peñuela, Albert, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, and others, 'Semantic Technologies for Historical Research: A Survey', *Semantic Web — Interoperability, Usability, Applicability*, 6.6 (2015), 539–64 <https://doi.org/10.3233/SW-140158>

Murano, Francesca, and Achille Felicetti, *Definition of the CRMtex*, 2020 <http://www.cidoc-crm.org/crmtex/ModelVersion/version-1.0-0>

Oberreither, Bernhard, *Intertextual Relationships Ontology (INTRO)*, 2020 <https://w3id.org/lso/intro/currentbeta#>

Peroni, Silvio, and David Shotton, 'FaBiO and CiTO - Ontologies for Describing Bibliographic Resources and Citations', *Journal of Web Semantics*, 17 (2012), 33–43 <https://doi.org/10.1016/j.websem.2012.08.001>

Sahle, Patrick, *A Catalog of Digital Scholarly Editions*, 2008- <https://www.digitale-edition.de/>

Schmidt, Desmond, and Robert Colomb, 'A Data Structure for Representing Multi-Version Texts Online', *International Journal of Human-Computer Studies*, 67.6 (2009), 497–514 <https://doi.org/10.1016/j.ijhcs.2009.02.001>

Sharing Ancient Wisdoms Ontology (SAWS), 2012 <https://ancientwisdoms.ac.uk/method/ontology/>

Tittel, Sabine, Helena Bermúdez-Sabel, and Christian Chiarcos, 'Using RDFa to Link Text and Dictionary Data for Medieval French', in *Proceedings of the 6th Workshop on Linked Data in Linguistics* (presented at the LDL-2018, Myazaki, Japan, 2018) <http://lrec-conf.org/workshops/lrec2018/W23/pdf/10_W23.pdf>

Tomasi, Francesca, 'Digital Editions Between Embedded Markup and External Representation. A Case Study: Vespasiano da Bisticci's Letters', in *Dall'informatica umanistica lle culture digitali*, ed. by Fabio Ciotti and Gianfranco Crupi, Quaderni Digilab, 2 (Sapienza Università Editrice, 2012), 201–18 <http://www.editricesapienza.it/sites/default/files/Quad_DigiLab_Informatica_Umanistica_Culture_Digitali.pdf>

———, ed., *Vespasiano da Bisticci, Lettere. A Knowledge Site* (Università di Bologna: /DH.arc, 2020) <https://web.archive.org/save/https://projects.dharc.unibo.it/vespasiano/>

———, ed., *Vespasiano da Bisticci, Lettere. A Semantic Digital Edition*, 2013 <https://doi.org/10.6092/unibo/vespasianodabisticciletters>

Vogeler, Georg, 'Digital Edition of Archival Material - Machine Access to the Content: On the Role of Semantic Web Technologies in Digital Scholarly Editions', in *Digitizing Medieval*

Sources ? L'édition en ligne de documents d'archives médiévaux, ed. by Christelle Balouzat-Loubet (Turnhout: Brepols, 2020), 37–56 <https://doi.org/10.1484/M.ARTEM-EB.5.117327>

Infrastructures and Technologies

Connecting TEI Content Into an Ontology of the Editorial Domain

Peter Boot, Marijn Koolen

Abstract

We argued elsewhere that, in order to support interoperable annotations, editions should provide machine-readable identifiers for text and text fragments, as well as information about the text fragments' type and structure. That is to say, they should be embedded in a Linked Open Data context that facilitates interchange and interpretation of annotation. In this article, assuming a TEI context, we consider the practical question of how the relevant RDF triples are to be derived. How is the edition to know which URIs are to be assigned to which elements in the XML hierarchy, and what are the relevant classes and properties? We discuss different options. Our preference is to generate the relevant triples upon ingestion of the XML file in a version control system and then to store the triples in the TEI `xenodata` element. We briefly consider situations in which cases the fine-grained annotation that we want to facilitate might be appropriate, or not.

1 Introduction

Users of online digital editions have a need for annotation support to contribute explanatory material that complements what is already available in the edition itself, for purposes of private study or for publication in conjunction with a scholarly article (Boot 2009; Robinson 2005; Siemens et al. 2012). One challenge for browser-based annotation support is anchoring the annotation to a specific location in the digital edition, as the browser typically only has an HTML representation of the edition that describes the page layout.

Recently, we proposed an ontology-based approach to describe digital editions and their components to a browser-based annotation tool, so that it understands and can reason about what it is annotating (Boot et al. 2017; Boot & Koolen 2018). While the ontology was formulated in order to support principled and robust targeting of annotations, it should have much wider applicability to ensure interoperability between the edition and other information items. As far as we know, no other annotation platform has developed similar functionality.[1]

[1] For example, Hypothesis (https://web.hypothes.is/) uses only page-level metadata in managing annota-

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 9–29.

In our approach, the edition server will describe the components of the edition to annotation tools (as well as to other clients). This description will be in terms of an ontology for the editorial domain, for which we have proposed a draft. The question that we discuss in this paper is: on the basis of which information is the edition application going to provide this information? How does it know to which classes certain text fragments belong, which URIs to assign, and how to name the property that connects, say, a chapter and its paragraphs? Assuming a Text Encoding Initiative (TEI) context, where the user interface of the edition is generated from (a) TEI source(s), the natural answer to these questions would be to store the relevant information in, or at least with, these TEI sources. This implies the need for TEI files to refer to the Annotation Ontology: the fragments of the edited text have to be assigned URIs; they have to be assigned a class; and, their mutual relationships have to be defined in terms of the properties described in the Ontology. In other words, we need to overlay the graph model describing the edition and its content on the (hierarchical) TEI XML.

The aim of this paper is not to give a final answer to the question of how and where to store the source information for the linked data necessary to facilitate interoperable annotation on a digital edition; rather we discuss a number of options, with their advantages and disadvantages.

What we discuss here is not necessarily in general the best way for embedding RDF in TEI XML. One thing to keep in mind is that, in our case, the RDF can be deduced from the XML structure, based on rules. This situation is fundamentally different from e.g., that of the SAWS project (see below), where the RDF triples represent new knowledge.

2 The Annotation Ontology

The Annotation Ontology is based on concepts from the FRBRoo ontology (IFLA 2015), which combines concepts from FRBR (Functional Requirements for Bibliographic Records) and CIDOC-CRM (The Conceptual Reference Model of the International Committee of Documentation, CIDOC 2006) to describe a.o. manuscripts and their editions (Le Bœuf 2012). With the Ontology, it is possible to describe the abstract work and its multiple representations as an RDF graph.

For a fuller discussion of the Ontology, we refer to Boot and Koolen 2018. The Ontology (See Figures 1 and 2) contains two important dimensions. First, a distinction between two domains: the editable and the edition domains. And, second, a distinction between works and documents.

tions. Pundit (https://thepund.it/) allowed web pages to specify URIs corresponding to HTML elements, but there was no possibility to assign classes, properties and relations to these URIs.

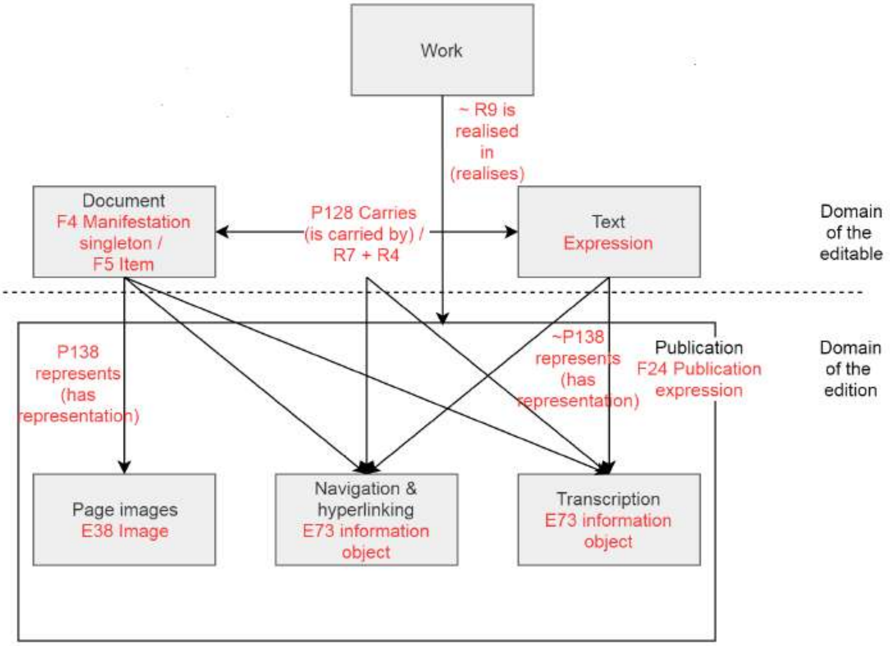

Work
~ R9 is realised in (realises)
Document
F4 Manifestation singleton / F5 Item
P128 Carries (is carried by) / R7 + R4
Text
Expression
Domain of the editable
P138 represents (has representation)
~P138 represents (has representation)
Publication
F24 Publication expression
Domain of the edition
Page images
E38 Image
Navigation & hyperlinking
E73 information object
Transcription
E73 information object

Figure 1. The Annotation Ontology extends the FRBRoo and CIDOC-CRM ontologies. The concepts in black are concepts in our annotation ontology. The concepts in grey are their (rough) equivalents in FRBRoo and CIDOC-CRM.

First, the editable domain has concepts for objects that are edited, such as documents and works, while the edition domain has concepts for the outcome of the editing process, such as digital texts and images. The two domains are connected. Within each domain, the classes are subclasses of a main class, of respectively editable things (EditableThing, left side of Figure 2) and edition things, (EditionThing, right side of Figure 2). As they are all potential targets for annotation, they are all subsumed under a main class of annotatable things (AnnotatableThing).

Second, the Work is an abstract intellectual and creative entity instantiated in one or more Documents, or text bearers. Both Works and Documents can have parts. Following Robinson (2017), we also describe what we call Positioned Text Fragments (PTF): intersections of the Work and Document hierarchy, such as the part of a poem that appears on a certain manuscript page.

What we discuss here is a generic ontology that should be suitable for most digital editions. We anticipate that edition projects with special needs and the required

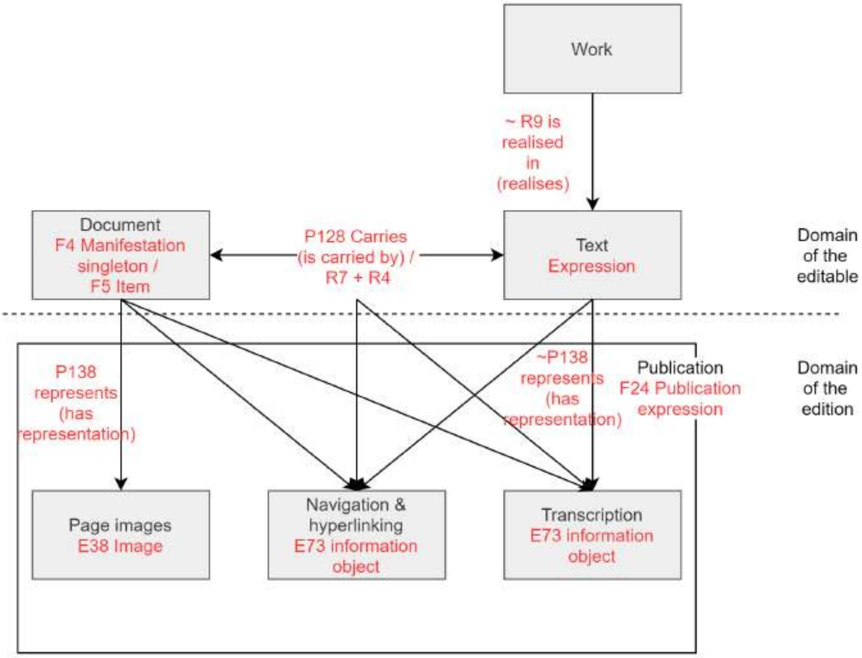

Work
~ R9 is realised in (realises)
Document
F4 Manifestation singleton /
F5 Item
P128 Carries (is carried by) / R7 + R4
Text
Expression
Domain of the editable
P138 represents (has representation)
~P138 represents (has representation)
Publication
F24 Publication expression
Domain of the edition
Page images
E38 Image
Navigation & hyperlinking
E73 information object
Transcription
E73 information object

Figure 2. The Annotation Ontology of the editable and edition domains.

expertise in ontology engineering will extend the generic ontology for their purposes. For an edition of medieval manuscripts, this specialized ontology might contain classes representing manuscript lines.

To illustrate how this Ontology can be used to describe digital editions, we use the digital edition of the correspondence of Vincent van Gogh (Jansen et al. 2009). Van Gogh's correspondence is a good example of the type of material that would benefit from interoperable annotation. Not only does the modern edition, as we will see, contain multiple representations of the same letter; the letters are also present in other forms or translations, on a number of other platforms (Douma n.d.; Koninklijke Bibliotheek 2016; Van Gogh 2019). In the modern edition, the abstract letter (middle column in Figure 3) is an instance of the class `vg:Letter`, which is a subclass of the FRBRoo class `F1 Work`.[2] It is a complex work consisting of multiple paragraphs, with

[2] For the Van Gogh edition, we use a specialized version of the ontology, as mentioned above; classes from the specialized ontology are prefixed by `vg`; for the general ontology we use the `hi` prefix.

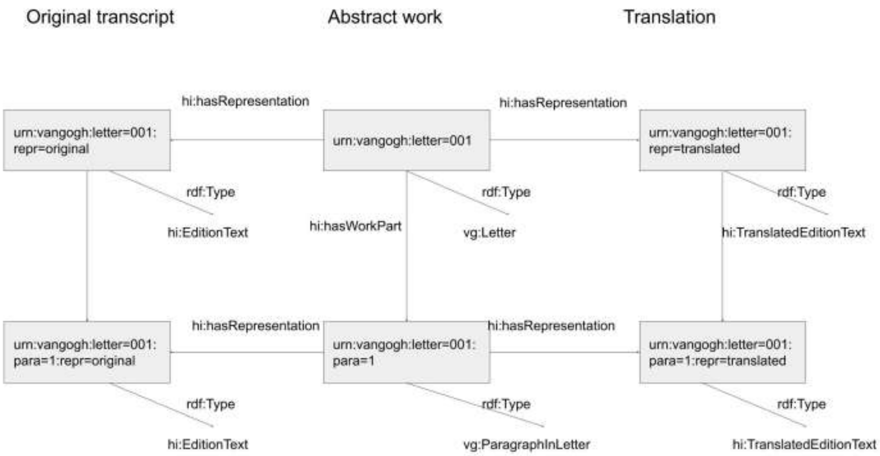


Figure 3. An example of the Annotation Ontology describing a letter of Vincent van Gogh as abstract work in the editable domain (middle), and its representations in the edition domain in the form of a reading text in Dutch (left), and a translation in English (right).

each individual paragraph being an instance of Paragraph which is a subclass of (Part of) Work. The abstract letter and its paragraphs are related to two representations, namely a Dutch text (left-hand column in Figure 3) and an English translation (right-hand column in Figure 3). Each instance has a unique identifier that can be the target of an annotation. To make annotations interoperable, the identifiers should be generated according to a set of guidelines, so that different online editions, whether based on the same TEI files or not, use the same identifiers. In the next section we discuss how such identifiers could be generated from the TEI files.

3 Connecting TEI and the Annotation Ontology

In this section we will discuss a number of approaches to storing the RDF that is required for attaching the edition to the annotation ontology. For each approach, we look at how that approach attaches the URIs to the relevant objects and how it stores the triples defining the relations between the URIs. We discuss two cases: first, a case where we distinguish between a work and its (multiple) representations in the edition, and second, the familiar issue of overlapping page and textual hierarchies. For both cases, we use, as an example, a letter by Vincent van Gogh. For all of the approaches, XML and XSLT samples are available in GitHub (Koolen et al. 2019a).

We are aware that these are relatively complex cases. There may be textual traditions where all annotation is, say, by line number in a single manuscript. In such cases, the required number of triples to describe the relevant context might be less daunting than in the cases we are about to discuss.

3.1 Criteria

The solution that we seek should satisfy a number of criteria: (i) it should be minimally obtrusive; (ii) maximally explicit; (iii) minimally redundant; and (iv) maximally generic and flexible. Preferably, (v) it should work without extending the TEI.

Ad (i): XML is sometimes seen as verbose and distracting (Bambaci et al. 2018). It is important that human readability and manageability of the XML is not impaired by computer code that is meaningless for the textual researcher.

Ad (ii): As much as possible, all information necessary for creating the edition should be contained in the XML file. If it is not included in the file, the file should unambiguously indicate how additional information is to be obtained. This is important for preservation purposes: an archived TEI file should contain all of the information required to create an edition.

Ad (iii): Information should not be unnecessarily duplicated

Ad (iv): The solutions we propose should not restrict editors to using TEI in a way that would otherwise be unsuitable for their projects. In fact, none of the approaches that we test create serious constraints.

Ad (v): Ideally, the solution should work in TEI as is. But, we should not be afraid of extending it where necessary.

3.2 One Work, Multiple Representations

The situation that we discuss here is the one depicted in Figure 3: one work (a letter), for which two texts are available: a version in the original language, and a translated version. We assume a situation where the annotatable units are the letter/text as a whole, and its individual paragraphs. The required RDF identifies the work, texts and paragraphs, creates `hasRepresentation` properties between work and texts (paragraphs) and describes the hierarchies between the works, texts and their paragraphs.

Apart from practical and convenience issues (readability), the essential question here is, how to make it possible for a single XML hierarchy to be associated with two conceptual hierarchies: the hierarchy of the work and the hierarchy of the work's representation.

GRDDL, Private URI Schemes and Dedicated URI Attribute

The GRDDL approach is based on the fact that for all triples, the necessary information is already contained in the XML file. Rather than including redundant information in the XML, this approach takes advantage of the GRDDL standard to associate an XML file with a program that derives an RDF representation of its contents (Conolly 2007)[3]. For defining the URIs, we create a dedicated attribute (here called `ontRef`). To keep the URIs short, we use the TEI facility of a private URI scheme: a prefix that is associated with a `prefixDef` in the TEI header which defines a pattern replacement to create a URI out of a brief pattern (TEI Consortium 2019, section 16.2.3). To associate two URIs (for work and text) with a single XML element, we create two prefix definitions for the same prefix.[4]
In practice, this approach might look like this:

Association between the TEI file and a GRDDL script:

```
<TEI xmlns:vg="http://www.vangoghletters.org/ns/"
     xmlns="http://www.tei-c.org/ns/1.0"
     xmlns:grddl="http://www.w3.org/2003/g/data-view#"
     grddl:transformation="grddl.xsl">
```

`prefixDefs` at the letter (`div`) level, defined in the TEI header:

```
<prefixDef ident="letter" matchPattern="([a-z]+),([0-9]+)"
           replacementPattern="urn:vangogh:letter=$2">
  <p>Associates letter div with uri of letter as work</p>
</prefixDef>
<prefixDef ident="letter" matchPattern="([a-z]+),([0-9]+)"
           replacementPattern="urn:vangogh:letter=$2:repr=$1">
  <p>Associates letter div with uri of this edition of the letter</p>
</prefixDef>
```

`ontRef` attribute on `div`s holding original and translated text:

```
<div type="original" ontRef="letter:original,1">
<div type="translation" ontRef="letter:translated,1">
```

A fragment of GRDDL-generated RDF in turtle format:

```
<urn:vangogh:letter=1> rdf:type vg:Letter.
<urn:vangogh:letter=1:repr=original> rdf:type hi:EditionText.
<urn:vangogh:letter=1> hi:hasRepresentation
<urn:vangogh:letter=1:repr=original>.
<urn:vangogh:letter=1:para=1> rdf:type vg:ParagraphInLetter.
<urn:vangogh:letter=1:para=1:repr=original> rdf:type hi:EditionText.
<urn:vangogh:letter=1:para=1> hi:hasRepresentation
<urn:vangogh:letter=1:para=1:repr=original>.
<urn:vangogh:letter=1> hi:hasWorkPart <urn:vangogh:letter=1:para=1>.
<urn:vangogh:letter=1:repr=original> hi:hasTextPart
<urn:vangogh:letter=1:para=1:repr=original>.
```

[3] A similar solution is discussed in (Daquino et al. 2019).

[4] An option discussed in the same section of the *TEI P5 Guidelines*.

At the top of the document, the reference (in the GRDDL namespace) to `grddl.xsl` points to the stylesheet to be executed in order to create an RDF representation of the document. This stylesheet, when processing the XML file, encounters the `ontRef` attribute on the `divs` holding the original language-version. It recognizes its value `letter:original,1` as a private URI scheme reference with the prefix `letter`. It searches for corresponding prefixDef elements and encounters two of these. It executes the corresponding match-and-replace operations, and creates these two URIs: `urn:vangogh:letter=1/` and `urn:vangogh:letter=1/repr=original/`. The logic for defining the triples (i.e., to assign the URIs to classes and to describe their properties) is contained in the stylesheet.

Advantages and disadvantages

We discuss advantages and disadvantages in terms of the criteria that we mentioned above.

Minimally obtrusive Referring to a GRDDL stylesheet does not affect the readability of the XML. Including `ontRef` attributes for each addressable element to hold an abbreviated pointer does affect readability, even though it could be much worse (if we included full URI's, for example).

Maximally explicit To generate the triples, the GRDDL stylesheet needs to be run. This implies that we need to be able to run an XSLT processor before the full information is available. The chances of being able, without significant effort, to run an XSLT processor in, say, fifty years' time, are slim. From that perspective this is a very weak solution. Indeed, GRDDL itself "is still a mere theoretical specification" (Grüntgens & Scharde 2016). A positive aspect is that usage of the dedicated `ontRef` attribute to refer to the URI is more explicit than using a general-purpose attribute such as `corresp`. A less than desirable aspect, however, is that the software has no way of knowing which of the `prefixDefs` produces the work-URI and which the text-URI.

Minimally redundant The triples are only generated as needed and therefore raise no redundancy concerns. The abbreviated pointers also avoid redundancy. However, the prefix definitions are included in each of the XML files, which, for the van Gogh edition, means more than 900 times. They could be included through `XInclude`.

No TEI extension The solution requires two new attributes: `grddl:transformation` and `ontRef`.

GRDDL, No URI Representation in TEI File

This approach, like the preceding one, uses GRDDL to deduce the necessary triples rather than include their representation in the TEI file. It goes one step further as it also leaves the URI definition out of the TEI. Instead, the TEI file contains `xml:id` attributes, and the GRDDL stylesheet generates triples that reference the `xml:id` attributes, as follows:

`div` elements with `xml:id` attributes:

```
<div type="original" xml:id="letter.original.1">
<div type="translation" xml:id="letter.translated.1">
```

Generated triples included now (beyond the ones from the previous approach):

```
<urn:vangogh:letter=1> hi:refersTo file:/.../let001.xml#letter.original.1.
<urn:vangogh:letter=1:repr=original> hi:refersTo file:/.../let001.xml#letter.
     original.1.
```

In this case, the stylesheet handles the generation of the URIs. The XML no longer needs to contain an `ontRef` attribute and `prefixDefs` to define the logic of the URI generation. The XML must provide `xml:id` attributes for the elements that are to be attached to a URI.

Advantages and disadvantages

Minimally obtrusive Referring to a GRDDL stylesheet does not affect the readability of the XML. The relevant elements need to carry `xml:id` attributes, but this is good practice anyway.

Maximally explicit This approach is even less explicit than the previous one, as we have now removed the URI definition (through the `prefixDefs`) from the XML into the GRDDL stylesheet.

Minimally redundant An improvement with respect to the previous approach, as the `prefixDefs` no longer need to be defined in each TEI file.

No TEI extension The approach requires one new attribute: `grddl:transformation`.

RDF in `xenodata`, URIs Generated from `xml:id`

This approach does not use GRDDL. The information necessary to deduce triples and URI has to be included in the TEI/XML file. Here we store that information in the xenodata element in the TEI header, introduced expressly to contain non-TEI (but still XML) metadata. The triples will therefore be included as RDF/XML. In this approach, we store the generated URIs on the relevant elements in a newly defined `ontRef` attribute.

In any realistic scenario, this still implies the need for a stylesheet or other program to generate the triples. Manually typing the RDF would be prohibitively time-consuming and error-prone. The difference with the GRDDL approach is twofold: (1) the output here is an enriched XML file, rather than (in the GRDDL case) an external

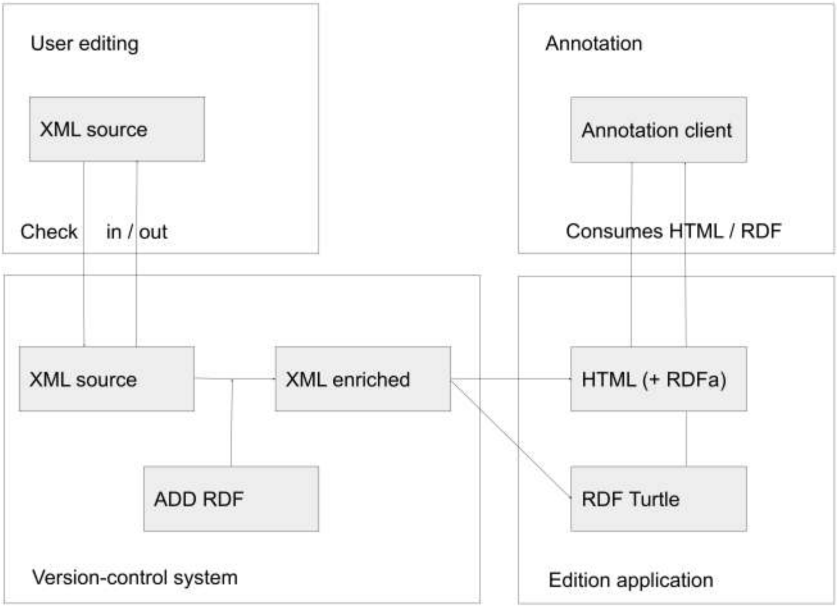


Figure 4. Overview of the editing and annotation workflow.

RDF file; and (2) here, the encoder is responsible for execution of the stylesheet before the TEI is published, while, in the GRDDL case, the stylesheet is (theoretically) executed by the consumer of the XML file after publication. This also implies that, for the benefit of the annotation tool (and possibly other RDF processing tools), the edition server will lift the triples out of the TEI environment, into a format that the annotation tool will be able to understand.

We envisage a situation where an editor works on an XML file without the RDF information. Upon ingestion in a version control system, the system executes a post-editing program that creates a file enriched with RDF. The editor will never need to see the enriched file, but keeps working on the original file. This is illustrated in Figure 4.

A fragment of the generated `xenodata` element might look like this:

```
<tei:xenodata xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
              xmlns:hi="http://boot.huygens.knaw.nl/vgdemo1/
                                 editionannotationontology.ttl#">
```

```
 <rdf:RDF>
  <rdf:Description about="urn:vangogh:letter=1" rdf:type="&vg;Letter"/>
  <rdf:Description about="urn:vangogh:letter=1:repr=original"
                   rdf:type="&hi;EditionText"/>
  <rdf:Description about="urn:vangogh:letter=1"
                   hi:hasRepresentation="urn:vangogh:letter=1:repr=original"/>
... ...
 </rdf:RDF>
</tei:xenodata>
```

As usual in XML/RDF, here we use XML entities (`&hi;`, `&vg;`) as abbreviations, because namespaces don't work in xml attributes. The entities are defined at the top of the document:

```
<!DOCTYPE any [
  <!ENTITY hi 'http://....../.../editionannotationontology.ttl#'>
  <!ENTITY vg 'http://....../.../vangoghannotationontology.ttl#'>
]>
```

The generated URIs for the letter's components are stored in their `ontRef` attribute (two URIs! One for the work, one for the edited text):

```
<div type="original" xml:id="letter.original.1" ontRef="urn:vangogh:letter=1
     urn:vangogh:letter=1:repr=original">
  <ab xml:id="para.original.1.1" ontRef="urn:vangogh:letter=1:para=1
      urn:vangogh:letter=1:para=1:repr=original">
[content of para]
  </ab>
...
</div>
```

Advantages and disadvantages

Minimally obtrusive If the RDF enrichment can indeed be handled as a post-editing process, where there is no need for the editor to work with the enriched file, the XML remains maximally readable for the editor.

Maximally explicit All RDF is contained in the enriched XML. In that sense, the approach is fully explicit. It might be considered a disadvantage that the RDF is not easily accessible to general-purpose RDF tools; however, the conversion to, say, a turtle file, is trivial.

Minimally redundant The choice for a post-editing approach, where the original XML and the enriched XML will be maintained separately for an indefinite period of time, entails some redundancy. However, as the enriched file is automatically created upon check-in in the version control system, this need not present any problem.

No TEI extension The approach requires adding an `ontRef` attribute to the TEI schema.

Triple Representation on Relation Element, URIs Generated From `xml:id`

Here we follow the choice made by the SAWS project (Jordanou et al. 2012). In this scenario, the relation element, originally meant to describe relations between persons, is used to carry any sort of triple. The `@active` attribute points to the subject of the triple, the `@ref` contains the property and the `@passive` attribute contains the object. The SAWS project used the `@active` and `@passive` attributes to refer to the `xml:id` attributes of the corresponding xml elements. In our case, this will not work, because we need to refer to the xml elements in their double capacity of work and text. We need to express, for instance, that the text is a representation of the work. On the relation element we, therefore, use URIs rather than just pointers to `xml:id`. As we assume the post-editing scenario outlined above, using full URIs presents no legibility problems.

Some of the generated relation elements will look like this:

```
<tei:listRelation xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <tei:relation active="urn:vangogh:letter=1" ref="rdf:type"
                passive="http://.../vangoghannotationontology.ttl#vg:Letter"/>
  <tei:relation active="urn:vangogh:letter=1:repr=original" ref="rdf:type"
                passive="http://.../editionannotationontology.ttl#hi:EditionText"
                    />
  <tei:relation active="urn:vangogh:letter=1"
                ref="http://.../editionannotationontology.ttl#
                    hi:hasRepresentation"
                passive="urn:vangogh:letter=1:repr=original"/>
...
</tei:listRelation>
```

Note that this is essentially the same solution as the xenodata approach, expressed in a different syntax. The advantages and disadvantages are similar.

Triples as RDFa, URIs Generated from `xml:id`

RDFa was designed with the purpose of embedding RDF into hierarchical languages such as HTML and XML (Herman et al. 2015). The idea is that the information that is already contained in the document can be labeled with semantic information using a number of extra attributes, the most important of which are about, `resource`, `typeof`, and `property`. An RDFa processor can then extract RDF from the document by combining the structure and content of the document with the RDFa labels. The approach was successfully used in Tittel et al. (2018).

RDFa is mostly used for expressing hierarchical structures, for example an agenda with agenda items. A very common structure is that a URI is assigned to a high-level element (say, a `div` holding the agenda) using the `about` attribute; underlying elements (say agenda items in `p` elements) are assigned a URI through a `resource` attribute; and the relation between the top and the lower level element is expressed using a property attribute. For each element, the `typeof` attribute defines the class that the attribute belongs to.

A single hierarchy, for example the relation between the work and its paragraphs, could be easily expressed using RDFa:

```
<div type="original" xml:id="letter.original.1" about="urn:vangogh:letter=1"
     typeof="vg:Letter">
  <ab xml:id="para.original.1.1" about="urn:vangogh:letter=1:para=1"
      typeof="vg:ParagraphInLetter" property="hi:hasWorkPart">
    <!-- content of para -->
  </ab>
  <!-- rest of paragraphs -->
</div>
```

The hierarchy of the XML translates naturally into the triple describing the relation between work and paragraph using the property `hi:hasWorkPart`.

However, in the case of our double hierarchy this no longer works. We have either to duplicate the entire hierarchy elsewhere in the XML document, or to create the triples representing the hierarchy manually by adding empty `seg` elements.
A fragment of the RDFa-enhanced XML could look like this:

```
<div type="original" xml:id="letter.original.1" about="urn:vangogh:letter=1"
     typeof="vg:Letter">
  <tei:seg resource="urn:vangogh:letter=1:para=1" property="hi:hasWorkPart"/>
  <!-- similar segs for other work paras -->
  <tei:div resource="urn:vangogh:letter=1:repr=original" property="
       hi:hasRepresentation"
           typeof="hi:EditionText">
    <tei:seg resource="urn:vangogh:letter=1:para=1:repr=original" property="
         hi:hasTextPart"/>
    <!-- similar segs for other text paras -->
    <ab xml:id="para.original.1.1" about="urn:vangogh:letter=1:para=1" typeof="
         vg:ParagraphInLetter">
      <tei:seg resource="urn:vangogh:letter=1:para=1:repr=original"
               property="hi:hasRepresentation" typeof="hi:EditionText">
        <!-- content of para -->
      </tei:seg>
    </ab>
    <!-- rest of paragraphs -->
</tei:div>
</div>
```

The outer `div` defines the letter (work) URI. The first empty `seg` element inside it relates the letter to its first paragraph. To simplify, we only show the first of these. The inner `div` defines the text URI. Here too, there is an empty `seg` to relate it to its (first) paragraph. The paragraph (`ab` element) carries the work paragraph URI. Inside, there is a `seg`-element that carries the text paragraph URI.

Advantages and disadvantages

Minimally obtrusive As this is again a post-editing process, the original XML remains readable for the editor. However, the enriched RDFa-XML has a convoluted structure and becomes near unreadable.

Maximally explicit As in the xenodata and relation approaches, all RDF is contained in the enriched XML. Here too, it can be trivially extracted for use by general-purpose RDF tools.

Minimally redundant Manageable redundancy, as above.

No TEI extension The approach requires adding the RDFa attributes to the TEI schema.

3.3 Text and Page Hierarchies

The second example that we discuss is concerned with the annotation of fragments of text in a double hierarchy: the logical hierarchy of text and text parts and the physical hierarchy of pages and lines. It should be possible to reuse an annotation on a certain manuscript line, both in the context of other annotations on that page and in the context of other annotations on the relevant paragraph. Robinson (2017) has devised the DET (Document Entity Text) addressing scheme that allows pointing at intersections of the logical and physical hierarchies, which we have called Positioned Text Fragments (PTFs). A PTF might be a part of a paragraph as it appears on one line in one manuscript; it might also be a part of a work as realized on one manuscript page.

In the example that we discuss here, we look at the intersections of pages and paragraphs. For a case of overlap between those hierarchies, we switch our attention to letter 4. Paragraph 4 is split over two pages. The situation that we want to represent is shown in Figure 5.

The main issue in representing this graph in the XML source is that the positioned text fragments are not (usually) represented in the XML, as such: they exist only implicitly, as the intersection between the logical hierarchy and the (milestone-based) physical hierarchy.

The XML of paragraph 4 looks like this (simplified):

```
<ab>
  How sorry I am about Uncle Hein. I sincerely hope he'll get better, but Theo, I
  fear he won't. Last summer he was still <pb/> so full of ambition, and had so
  many plans and told me that business was going so well. It is indeed sad.
</ab>
```

Representing the paragraph-page PTFs would turn this into something like:

```
<ab>
  <ptf>How sorry I am about Uncle Hein. I sincerely hope he'll get better, but
       Theo, I fear he won't. Last summer he was still</ptf>
  <pb/>
  <ptf> so full of ambition, and had so many plans and told me that business was
       going so well. It is indeed sad.</ptf>
</ab>
```

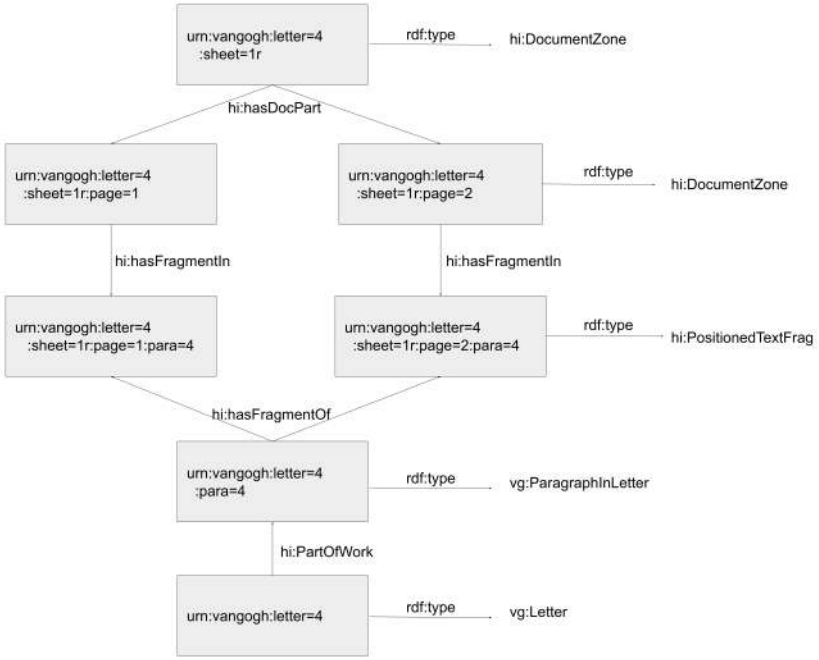

urn:vangogh:letter=4 :sheet=1r
rdf:type
hi:DocumentZone
hi:hasDocPart
urn:vangogh:letter=4 :sheet=1r:page=1
urn:vangogh:letter=4 :sheet=1r:page=2
rdf:type
hi:DocumentZone
hi:hasFragmentIn
hi:hasFragmentIn
urn:vangogh:letter=4 :sheet=1r:page=1:para=4
urn:vangogh:letter=4 :sheet=1r:page=2:para=4
rdf:type
hi:PositionedTextFrag
hi:hasFragmentOf
urn:vangogh:letter=4 :para=4
rdf:type
vg:ParagraphInLetter
hi:PartOfWork
urn:vangogh:letter=4
rdf:type
vg:Letter

Figure 5. RDF representation of two positioned text fragments in the text and page hierarchies.

If there were an italicized phrase around the page break, we would have to split the PTFs. Using the n-attribute to connect what should logically be considered as a single PTF, we get:

```
<ab>
  <ptf n="1">How sorry I am about Uncle Hein. I sincerely hope he'll get better,
             but Theo, I fear he won't. </ptf>
  <emph>
    <ptf n="1">Last summer he was still </ptf>
    <pb/>
    <ptf n="2"> so full of ambition</ptf>
  </emph>
  <ptf n="2">, and had so many plans and told me that business was going so well.
             It is indeed sad.</ptf>
</ab>
```

This would be clearly impracticable in a file that should be usable for editors. This seems to imply that we have the choice between a GRDDL-like solution, where a client dynamically generates RDF at runtime, or a post-editing script, as discussed in some of the other approaches above.

However, the GRDDL solution will not actually work. A GRDDL script generates RDF that other programs can consume, but it leaves the XML unchanged. The generated RDF will use URIs to refer to PTFs, but in the XML there will be nothing that corresponds to these URIs. To find out what an annotation of a PTF refers to, we'd have to inspect the logic of the GRDDL script. This implies that, in this case, a post-editing script is the only workable answer.

The result of the post-editing script would be an XML file enhanced with triples (e.g. in a `xenodata` element), with `pb` elements, `divs` and paragraphs provided with `ontRef` attributes, and `seg` elements with `type=ptf` to represent the positioned text fragments.

Apart from the work-level information about work and paragraphs that we've seen in the previous sections, the `xenodata` element would now also define the sheet of paper, as well as describe the pages on the sheet and the positioned text fragments. It would also describe the relations between these objects:

```
<tei:xenodata>
  <rdf:RDF>
    <rdf:Description
            about="urn:vangogh:letter=4:sheet=1r"
            rdf:type="http://....../.../editionannotationontology.ttl#
                 DocumentZone"/>
    <rdf:Description about="urn:vangogh:letter=4:sheet=1r"
                     hi:hasDocPart="urn:vangogh:letter=4:sheet=1r:page=1"/>
    <rdf:Description
            about="urn:vangogh:letter=4:sheet=1r:page=1"
            rdf:type="http://....../.../editionannotationontology.ttl#
                 DocumentZone"/>
    <rdf:Description
            about="urn:vangogh:letter=4:sheet=1r:page=1:para=4"
            rdf:type="http://....../.../editionannotationontology.ttl#
                 PositionedTextFrag"/>
    <rdf:Description about="urn:vangogh:letter=4:para=4"
                     hi:hasFragmentOf="urn:vangogh:letter=4:sheet=1r:page=1:para
                          =4"/>
    <rdf:Description about="urn:vangogh:letter=4:sheet=1r:page=1"
                     hi:hasFragmentIn="urn:vangogh:letter=4:sheet=1r:page=1:para
                          =4"/>
    <rdf:Description about="urn:vangogh:letter=4:sheet=1v:page=2"
                     hi:hasFragmentIn="urn:vangogh:letter=4:sheet=1v:page=2:para
                          =4"/>
    ... ...
  </rdf:RDF>
</tei:xenodata>
```

The triples would be about the URIs defined in the transcription:

```
<div type="original" xml:id="letter.original.4" ontRef="urn:vangogh:letter=4">
  <pb f="1r" n="1" xml:id="pb-orig-1r-1" facs="#zone-pb-1r-1"
      ontRef="urn:vangogh:letter=4:sheet=1r:page=1"/>
  <lb n="1" xml:id="l-1"/>
  <ab xml:id="para.original.4.1" ontRef="urn:vangogh:letter=4:para=1">
    <tei:seg type="ptf" ontRef="urn:vangogh:letter=4:sheet=1r:page=1:para=1">
      [content]
    </tei:seg>
  </ab>
...
</div>
```

The advantages and disadvantages of this approach are really the same as those of the xenodata approach, discussed above: the XML that the editor sees remains fully readable; the enriched XML contains a full RDF representation; the resulting redundancy is manageable; and, the only TEI extension that is needed is the addition of an `ontRef` attribute.

4 Conclusion and Discussion

There is no single best way to store the source for the RDF describing a text structure's embedding in the Linked Data world. We discussed a number of approaches for two different cases, but there are other cases. Combinations of approaches would also be possible. For example, to work around the double hierarchies in RDFa, we could define the work hierarchy under the `xenodata` element, and use RDFa to define the hierarchies of the representations (original text and translation). However, this would make it harder to lift the RDF triples out of the XML.

From the preceding discussion it appears that the GRDDL approach is weakest in terms of explicitness. The fate of the standard is unclear, and the archived TEI files will contain no representation of the triples or even, in our second GRDDL approach, of the URIs. GRDDL's strongest point, minimal redundancy (and therefore, user-friendly XML), can also be obtained if we implement a post-editing script in the version-control system, as used in the other approaches. Of these, the RDFa approach might be suitable for simple cases, where all annotation uses a single hierarchy. In other cases, the generated RDFa becomes unwieldy. The remaining approaches (using either the relation or the `xenodata` element) are very similar to one another. An advantage of the relation element might be that the triples can be stored near (one of) the elements that they refer to, while the xenodata approach collects all triples in the document's header. For our post-editing approach, this consideration is not really relevant. We prefer the xenodata approach because it remains closest to the RDF model.

Given that we propose to derive the triples upon ingestion in the version control system, our criterion of maximal explicitness would prescribe that the TEI files contain the name of the program that deduces the RDF statements. The version control system should use the program that the TEI file mentions. As yet we have not formulated a preference for how the program name should be stored in the TEI file. An obvious choice would be to include the information in an XML processing instruction. Another option would be to widen the scope of the existing TEI element `appInfo`, now used for recording past processing steps, to define also-required future processing steps. The latter choice would require a hook in the version control system that knows how to read TEI, and might be less flexible than using a processing instruction.

In this paper we have assumed that all URIs for the objects that we want to refer to are URNs; our objects are abstract information objects and not, for example, web pages. Our URNs are all persistent, as they should be. More specifically, we have made our URNs transparent: by constructing them as lists of numbered units (e.g., `'letter=4:sheet=1r:page=1:para=1'`) software as well as humans can understand them and act upon them. This is not strictly necessary. Projects that require opaque identifiers could choose to assign unique random identifiers to their objects. However, these URIs could no longer be generated on the basis of the XML's hierarchical structure alone and would have to be included in the XML.

5 Outlook

We have developed a prototype annotation tool that reads the RDFa describing the edition, as displayed in the browser, as well as any linked RDF descriptions of external components, and allows users to make various types of annotations (Koolen & Boot 2020). The software is open source and consists of a JavaScript client (Koolen et al. 2019a) that edition creators can easily incorporate in their online editions, and a server (Koolen et al. 2019b) that can store and retrieve the annotations. Edition creators can decide to run their own annotation servers or configure the client to communicate with other annotation servers. A next step for us will be to test this approach with a number of real editions.

Our approach to annotation creates the possibility for users to choose very specific targets for their annotation. One of the issues that we will have to confront during testing is whether this flexibility might not, paradoxically, decrease the interoperability of user annotations. The flexibility that we offer might be confusing to users, and lead them to choose inappropriate annotation targets. It is possible that the distinctions that FRBR makes are too subtle for the average user.[5] We do not expect this in itself to be a problem. After all, we design an annotation tool for scholarly use, and though FRBR terminology might be complex, distinctions such as the one between a work and a copy are the bread and butter of scholarship. Research about this issue is not really available due to the fact that annotation tools that support these distinctions are scarce. Most research about FRBR usability has been in the context of library catalogs. "FRBR-based displays do make sense to users and better support their tasks" write Zhang & Salaba (2012), and similar findings are reported in Kim (2015). Hunter & Gerber (2010) present, as far as we know, the only study of FRBR-based annotation.

[5] We thank Michele Pasin for this pertinent question.

They find that the FRBR ontology is, indeed, complex to many users. However, the tool they tested only supports annotation at the entry-level, not on text fragments. We really need more experimentation to understand how FRBR-based annotation will work in practice.

Meanwhile, within the domain of textual editing, our approach may be useful for other forms of interaction with the text, besides annotation. The URNs that we use are, not coincidentally, very similar to the URNs as used in the Canonical Text Services protocol (Blackwell & Smith 2014).[6] This could provide the basis for exchange of text fragments between applications and digital text collections. Beyond the domain of textual editing, a similar approach will be useful in other fields that have multiple representations of a single original object, such as in the field of video (consider representations in multiple resolutions, separate audio tracks or textual transcripts, or an entire news program vs. individual items). Even in photography, an annotation can be about the photographed object or about the picture itself. Distinguishing between object and representation, and embedding both in their context, is a key requisite for interoperable annotation and other applications.

Acknowledgements

This work is partly funded by the KNAW 'Vernieuwingsgelden' and by the NWO project *CLARIAH*.

Bibliography

Bambaci, Luigi, Federico Boschetti, and Riccardo Del Gratta, 'Qohelet Euporia: A Domain Specific Language to Annotate Multilingual Variant Readings' (presented at the 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), IEEE, 2018), 266–69

Blackwell, Chris, and Neel Smith, eds., 'Canonical Text Services Protocol Specification', version 2.0.rc.1, 2014 <http://cite-architecture.github.io/ctsurn_spec/>

Boot, Peter, *Mesotext. Digitised Emblems, Modelled Annotations and Humanities Scholarship* (PhD thesis, Utrecht University, 2009)

Boot, Peter, Ronald Haentjens Dekker, Marijn Koolen, and Liliana Melgar, 'Facilitating Fine-Grained Open Annotations of Scholarly Sources', *DH2017 Conference Abstracts*, 2017 <https://dh2017.adho.org/abstracts/198/198.pdf>

[6] Canonical text Services might evolve into Distributed Text Services (https://w3id.org/dts), the specification of which is currently a working draft. It proposes a slightly different reference structure, combining URNs with query parameters (see https://distributed-text-services.github.io/specifications/Navigation-Endpoint.html). We are following the developments of the DTS specification and will investigate how our approach can made interoperable with it.

Boot, Peter, and Marijn Koolen, ‘A FRBRoo-Based Annotation Ontology for Digital Editing’ (presented at the *‘Data in Digital Humanities’ - the European Association for Digital Humanities 2018* Conference, Galway, 2018) <https://eadh2018.exordo.com/files/papers/93/final_draft/A_FRBROO-based_annotation_ontology_for_digital_editing__Final_.pdf>

CIDOC, ‘CIDOC Documentation Standards Working Group, and CIDOC CRM SIG’, 2006– <http://www.cidoc-crm.org/>

Ciotti, Fabio, ‘A Formal Ontology for the Text Encoding Initiative’, *Umanistica Digitale*, 2.3 (2018) <https://doi.org/10.6092/issn.2532-8816/8174>

Connolly, Dan, ed., ‘Gleaning Resource Descriptions from Dialects of Languages (GRDDL). W3C Recommendation 11 September 2007’, 2007 <https://www.w3.org/TR/grddl/>

Daquino, Marilena, Francesca Giovannetti, and Francesca Tomasi, ‘Linked Data per le edizioni scientifiche digitali. Il workflow di pubblicazione dell’edizione semantica del quaderno di appunti di Paolo Bufalini’, *Umanistica Digitale*, 3.7 (2019) <https://doi.org/10.6092/issn.2532-8816/9091>

Douma, Michael, ed., *Van Gogh’s Letters. Unabridged and Annotated.* (Webexhibits.org, n.d.), <http://www.webexhibits.org/vangogh/>

Grüntgens, Max, and Torsten Schrade, ‘Data Repositories in the Humanities and the Semantic Web: Modelling, Linking, Visualising’, *WHiSe 2016 Humanities in the Semantic Web Proceedings of the 1st Workshop on Humanities in the Semantic Web co-located with 13th ESWC Conference 2016* (s.l.: CEU, 2016), 53–64

Hunter, Jane, and Anna Gerber, ‘The Aus-e-Lit Project: Advanced EResearch Services for Scholars of Australian Literature.’, in *VALA 2010.* Melbourne, Australia 2010 <http://vala.org.au/vala2010/papers2010/VALA2010_85_Hunter_Final.pdf>

Herman, Ivan, Ben Adida, Manu Sporny, and Mark Birbeck, ‘RDFa 1.1 Primer - Third Edition Rich Structured Data Markup for Web Documents’, W3C Working Group Note, 17 March 2015 <https://www.w3.org/TR/xhtml-rdfa-primer/>

IFLA, *Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*, 2.4 vols (The Hague, 2015)

Jansen, Leo, Hans Luijten, and Nienke Bakker, eds., *Vincent van Gogh: The Letters* (Amsterdam: Huygens ING, Van Gogh Museum, 2009), <http://vangoghletters.org/vg/>

Jordanous, Anna, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman, ‘Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web’, in *WIMS ’12. Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, ACM, 2012*, Art. no. 44 <https://doi.org/10.1145/2254129.2254184>

Kim, Sung-Min, ‘Towards Organizing and Retrieving Classical Music Based on Functional Requirements for Bibliographic Records (FRBR)’ (University of Pittsburgh, 2015)

Koninklijke Bibliotheek, *Geheugen van Nederland [Memory of the Netherlands]* (The Hague: Koninklijke Bibliotheek, 2016), <https://www.geheugenvannederland.nl/nl>

Koolen, Marijn, Jaap Blom, and Melgar Estrada Liliana, *Scholarly Web Annotation Client [Github Repo]*, 2019a, <https://github.com/CLARIAH/scholarly-web-annotation-client>

———, *Scholarly Web Annotation Server [Github Repo]*, 2019b <https://github.com/CLARIAH/scholarly-web-annotation-server>

Koolen, Marijn, and Peter Boot, 'Facilitating Reusable Third-Party Annotations in the Digital Edition', in *Annotation in Scholarly Editions and Research* (De Gruyter, 2020), 177–199

Le Bœuf, Patrick, 'Modeling Rare and Unique Documents: Using FRBRoo/CIDOC CRM', *Journal of Archival Organization*, 10.2 (2012), 96–106 <https://doi.org/10.1080/15332748.2012.709164>

Robinson, Peter, 'Some Principles for Making Collaborative Scholarly Editions in Digital Form', *DHQ: Digital Humanities Quarterly*, 11.2 (2017) <http://www.digitalhumanities.org/dhq/vol/11/2/000293/000293.html>

———, 'Where We Are with Electronic Scholarly Editions, and Where We Want to Be', *Jahrbuch Für Computerphilologie*, 5 (2005), 125–46 <http://www.computerphilologie.de/jg03/robinson.html>

Siemens, Ray, Meagan Timney, Cara Leitch, Corina Koolen, and Alex Garnett, '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), 445–61 <https://doi.org/10.1093/llc/fqs013>

TEI Consortium, *P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2019) <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>

Tittel, Sabine, Helena Bermúdez-Sabel, and Christian Chiarcos, 'Using RDFa to Link Text and Dictionary Data for Medieval French', in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, 7–12 <http://lrec-conf.org/workshops/lrec2018/W23/pdf/10_W23.pdf>

Van Gogh, Vincent, *Brieven aan zijn broeder (3 delen)* (DBNL, 2019), <https://www.dbnl.org/tekst/gogh006brie00_01/>

Zhang, Yin, and Athena Salaba, 'What Do Users Tell Us about FRBR-Based Catalogs?', *Cataloging & Classification Quarterly*, 50.5–7 (2012), 705–23 <https://doi.org/10.1080/01639374.2012.682000>

Towards Resolution Services for Text URIs

Hugh Cayless, Matteo Romanello

Abstract

In this paper we address the lack of fully resolvable URIs for texts and their citable units in the currently emerging Graph of Ancient World Data. We identify three main architectural components that are required to provide resolution services for text URIs: 1) a registry of text services; 2) an identifier resolution service; 3) a document metadata scheme, to represent the relations between texts in the registry, as well as between these texts and related external resources (e.g. library catalogues). After presenting some of the use cases a central registry providing resolvable URIs for texts would enable, we discuss in detail each component. We conclude by considering three examples where the proposed document metadata scheme is used to describe digital texts; this scheme contains a minimum yet extendable set of metadata that can be used to explore and aggregate texts coming from a network of distributed repositories.

1 Introduction

Over the last decade, many projects and institutions in the field of (digital) classics have embraced Linked Open Data as a philosophy of sharing and interconnecting resources about the ancient world (Elliott et al. 2014; Middle 2018), thus leading to the emergence of a Graph of Ancient World Data (Isaksen et al. 2014). The community contributing to the growth of this graph values the use of shared controlled vocabularies based on URIs to refer to *things*. Pleiades has produced an extensive set of URIs to identify ancient geographical locations and, more recently, Pelagios (Isaksen et al. 2014) has aggregated these, along with other gazetteers, and links to related resources. Something similar has been done by another two projects: SNAP (Bodard et al. 2017), and PeriodO (Rabinowitz et al. 2016), for ancient people, and chronological periodizations, respectively. Yet, what is still missing are fully resolvable URIs for texts and their citable units. Such URIs are of the highest importance for this emerging graph of data, considering that cited primary sources are an area where existing datasets often overlap.

Although there exist a number of platforms and identifier schemes for referring to text, there is no central registry for text identifiers and metadata (with a granularity down to the passage level). These kinds of identifiers are crucial in a Linked Data

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 31–44.

context, because they allow common reference to abstract works (e.g., the *Iliad*) and to specific editions of that work (e.g., the Venetus A manuscript of the *Iliad*, a digital transcription of the Venetus A, etc.). In addition, while there is adequate coverage for canonical Classical texts, the situation is much less well-defined for the many texts outside the canon (e.g., papyri and inscriptions).[1] Further, the existing platforms that deal with identifier resolution do not provide services for minting new identifiers or refactoring existing ones.

Given this gap in the current digital infrastructure, we propose some initial thoughts towards the creation of web services that are able to handle the resolution of URIs pointing to texts and their citable units.[2] These services are aimed at increasing the presence of textual resources in the emerging Graph of Ancient World Data, and at better connecting digital texts with other relevant resources available online.

2 Related Work

The resolution services for text URIs that we propose in this paper would not be even conceivable without the work that has been carried out in recent years in two directly-related research areas. Firstly, the development of Application Programming Interfaces (APIs) to facilitate the exchange of digital texts; and, secondly, the development of standard vocabularies for the semantic description of texts.

Concerning text APIs, some concrete solutions have been developed over the last decade to enable the exchange of structured texts over standard protocols. The Canonical Text Services (CTS) (The Canonical Text Service 2019; Smith 2009) was the first of such protocols to define an API as well as an identifier syntax – the so-called CTS URNs – to retrieve TEI-encoded texts. Unfortunately, some technological factors hindered CTS from becoming a widely-adopted standard for exchanging texts (see section 4.1), and have led to the development of a new API, the Distributed Text Services (DTS) (Distributed Text Services n.d.; Clérice 2017; Clérice et. al. 2017), which aspires at playing for text data, the role that the IIIF API has played for image data. In addition to specifications and working implementations of both CTS and DTS, there exists an ecosystem of tools to work with them, most notably the CapiTainS software suite (Clérice 2017), which provides support for both APIs.

However, the absence of a widely-adopted API to exchange structured texts has also led to the proliferation of ad hoc solutions, whose designs often bear striking similarities with that of CTS and DTS. These include, for example, the API developed

[1] Trismegistos (Trismegistos : An Interdisciplinary Portal of the Ancient World n.d.) assigns *TM numbers* to ancient documents, and can serve as a kind of gazetteer for some of these types of source text.

[2] The work presented in this paper is the result of Linked Texts, a working group recently funded by Mellon Foundation via the *Pelagios Commons* project, which brought together an international group of scholars and research developers with extensive experience in developing and/or working with APIs for text repositories.

for the Scholastic Commentaries and Texts Archive (SCTA) (Witt 2018), the SHINE API developed in the context of the *Research Infrastructure for the Study of Eurasia* project (RISE), or the API that exposes the textual data of the *School of Salamanca* project (Wagner 2019).

Besides initiatives focusing on text APIs, considerable efforts were made to devise vocabularies and ontologies for describing various aspects of texts, with applicability to the ancient world and beyond. The Linked Ancient World Data ontology (LAWD) (Cayless 2016) models some aspects of ancient texts, and, in particular, of text-bearing artefacts such as papyri and inscriptions, while aligning itself with other vocabularies like DC, OAC and CIDOC-CRM. The HuCit Knowledge Base (Romanello & Pasin 2017) builds upon existing ontologies (e.g. CIDOC-CRM and FRBRoo) to link together bibliographic resources about classical canonical texts such as Perseus Catalog and Classical World Knowledge Base (CWKB), and focuses on the citation structures and citable units of these texts. Finally, the SPAR family of ontologies (Peroni & Shotton 2018) provides a set of vocabularies to describe a wide range of aspects around publications, including the characterization of citing behaviors with the Citation Typing Ontology (CiTO) and the representation of bibliographic metadata with the FRBR-aligned Bibliographic Ontology (FaBiO).

Finally, there exist ontologies that could benefit from the availability of resolvable URIs for texts. In addition to the already mentioned CiTO that characterizes citational relationships between texts, ontologies like Sharing Ancient Wisdoms (SAWS) (SAWS The Ontology n.d.) or the Intertextual Relationship Ontology (INTRO) (Oberreither 2019) could be used to describe textual relations between various text passages. These ontologies already provide vocabularies that could be used to create semantic statements about portions of texts, if only there existed resolvable URIs for the text portions one may want to make statements about.

3 Use Cases

A central registry providing resolvable URIs for texts and text passages would enable a wide range of use cases:

1. Services and tools for annotation – be it manual or automatic – could use text URIs to support the enrichment of existing resources. Tools like Recogito (Isaksen et al. 2017) or INCEpTION (Klie et al. 2018) already enable users to annotate texts and images by using existing authority data available via a SPARQL API (e.g. Pleiades, Wikidata, etc.). Recogito users, for example, can annotate geographical places within texts by assigning to each place its corresponding identifier (URI) within a gazetteer (e.g. https://pleiades.stoa.org/places/579885 for Athens). Having resolvable URIs for texts will enable a similar usage scenario for textual materials.

2. Online publishers will use the data to create links between their digital publications and existing digital libraries/text services. Take, for example, the digital version of a book containing citations of primary sources, like the open access publications published by Harvard's Center for Hellenic Studies (Center for Hellenic Studies: Online Publications n.d.). The publisher may want to provide readers with click-through links to the full text of cited sources, while leaving them the freedom to pick a specific edition/translation, should multiple versions of a text or document be available. Also, creating links towards a resolver, rather than to specific digital libraries, can potentially reduce the risk of broken links (as the resolver can be updated to reflect URL changes of digital libraries).
3. Digital Libraries collections could be registered in our text registry, provided that they expose their text data in compliance with at least one of the APIs supported by the registry, with the effect of increasing the accessibility and discoverability of these collections.

4 Text Resolution Services

The main idea of a service able to create resolvable URIs for texts published online is to relieve the publishers of such texts from having to mint URIs for their texts and the citable units in them. If texts are made available via standard APIs and described by sufficient metadata, the creation and resolution of URIs can be delegated to a centralized service, thus making it easier to integrate textual resources into the LOD graph.

In order for such services to work, three main things are required: *identifiers* for texts and their citable units; *machine interfaces* (i.e. APIs) to access and retrieve texts over Web protocols; and *metadata* to describe texts, as well as the relations between these texts and other existing resources (e.g., authority control records).

We believe that three main components are needed to provide resolution services for text URIs:

1. *A registry of text services* to keep track of available text repositories against which text URIs can be resolved.
2. An *identifier resolution service* that can resolve a URI pointing to a section of a text or document (e.g., a line of a papyrus or inscription, or a chapter of a work in prose) to one or more places where a digital version of that text or document can be found.
3. A *document metadata scheme* is also needed in order to represent the relations between texts in the registry, as well as between these texts and related external resources (e.g., library catalogues).

In what follows we discuss the main issues related to each of these three components.

4.1 Text Registry Service

Supported APIs

CTS and DTS are the two main APIs that have emerged in recent years to enable the exchange of structured texts over standard protocols. We identified these two APIs as the most important standards that our text resolution services should support.

Unfortunately, some factors hindered CTS from becoming a widely adopted standard for exchanging texts:

1. its strong commitment to the specificities of canonical texts, which makes it unsuitable for non-canonical texts such as archival documents, papyri or inscriptions;
2. aspects of the design of the API which keep it from scaling to large repositories of texts;
3. the development of the CTS standard, which has been driven by a single project rather than a community.

While the use of the CTS protocol has been limited, the naming system it defines is powerful and has been broadly adopted. The DTS is a new standard whose specifications have recently been published as a first public draft. DTS was designed to overcome the limitations of CTS, discussed above, while remaining retro-compatible with existing CTS URNs. DTS presents a text API modelled as arbitrary collections, with functions for retrieving whole or partial texts, and for discovering the citation schemes for those texts.

Registry Implementation

A suitable model for implementing a text registry that is open to input from the community is w3id.org, where pull requests made to a GitHub repository make it possible to add new entries to the registry. Anyone should be able to add a service to this registry, provided that this service exposes its data via a CTS or DTS API. Moreover, since the DTS API advertises the citation structures and citable units of a text, it can be used to determine, programmatically, which resources URIs will have to be minted for.

4.2 Identifier Resolution Service

Text Identifiers

CTS URNs provide a handy way of creating text identifiers that carry with them information about the text itself and its hierarchy. Yet, there is no place where these CTS URN identifiers are published, no entities nor community regulating their creation, no central place where they are gathered.

Let us consider an example of how CTS URN work: Proclus' *Elements of Physics* (`urn:cts:greekLit:tlg4036.tlg006`). The Perseus Catalog contains bibliographic information about this text, including a link to an edition available in the Hathi Trust collection (*Perseus Catalog,* Institutio Physica n.d.). The SAWS project (Hedges et al. 2016) has added an edition of this text, and publishes it via its CTS endpoint. Their edition is identified by the URN `urn:cts:greekLit:tlg4036.tlg006.saws01`. Let us imagine now that new editions or translations of the same text are published via a DTS endpoint with compatible CTS URNs. How can all this information be integrated in such a way that it is available to a user looking up the identifier `urn:cts:greekLit:tlg4036.tlg006`?

Resolver or Catalog?

One of the main points of discussion at the Linked Texts working group meeting was how to handle the resolution of text URIs, and whether it was feasible to develop a Handle System, based on the one that is used for resolving Digital Object Identifiers, that could work with DTS and CTS APIs (Almas et al. 2018).

A Handle System works by querying a database for the submitted ID, and then, if a matching URL is found, redirecting the requester to that URL. The situation for DTS is more complex, partially because of the legacy of CTS identifiers, which are URNs. The problems begin with the lack of any central management of CTS IDs. Unlike DOIs, which consist of a namespace identifier and an item identifier, CTS URNs are more hierarchical and more meaningful at each level.

Take `urn:cts:greekLit:tlg0012.tlg001`, for example, which identifies the *Iliad* of Homer. Crucially, it identifies the *work* rather than any specific edition of it. Usefully, a CTS URN allows us to denote a passage either within the abstract work (`urn:cts:greekLit:tlg0012.tlg001.1.1-1.10`, i.e., *Iliad* 1, lines 1–10) or within a specific edition (`urn:cts:greekLit:tlg0012.tlg001.perseus-grc1.1.1-1.10`, i.e. the version of the *Iliad* hosted by the Perseus Project, book 1, lines 1–10).

This means there is a basis for identifying the common abstract notion of the first line of the *Iliad*, and comparing them across all available editions. But what should a resolver do, given a work ID? A possible answer would be for it to list available editions (and translations) of that work, behaving more like a catalog (or a regular DTS Collection endpoint) than a resolver. It gets worse when we consider that edition identifiers are uncontrolled, and that there is no reason that multiple repositories might not contain copies of `urn:cts:greekLit:tlg0012.tlg001.perseus-grc1` (and perhaps even different versions of that).

This complexity means that a *resolver* cannot work like a traditional handle server, because there are cases where there is not a one-to-one relationship between identifier and resource, for example, when a URI corresponds to a non-information resource (e.g., a work in the FRBR hierarchy), or when the URI can resolve to multiple copies of the same resource. In such cases, the resolver will deliver a list of documents (or document fragments), and the decision must be made by the client – be it a human or software agent – about which resource(s) to choose for a given identifier. In other words, despite the attractions of a handle resolution service, the resolver must instead behave in certain respects like a centralized catalog system.

4.3 Document Metadata Scheme

What sorts of information might we want to attach to a digital document? What information/metadata is needed in order to provide useful services on an aggregate of texts coming from a network of distributed repositories?

The kinds of document represented by DTS will usually be editions derived from one or more manuscript or print sources, and so there may be a natural division between data attaching to the source document(s), and to the edition or translation. This division might be handled in DTS by modelling the collections accordingly. Inscriptions, for example, will have information that refers to the primary source document itself: its find-spot, date, provenance, and so on. Then, there may be multiple published editions of that inscription, and these have their own data, including the editor and publication date. Inscriptions may be grouped in various ways, but one very typical way is to organize them geographically: a plausible DTS collection structure for inscriptions might be *location* → *sources* → *editions*, with source information attached to the source collection, and edition information to the edition. But this is only one possible organizational scheme. *Canonical* texts will tend to be organized according to author and work, where *author* refers to the creator of the original *work*, which is an abstraction – effectively an agreement that the sources all have the same ancestor. Many source texts may go into the creation of an edition (and not every edition will share the same set of sources). Thus, we may expect both source and edition information to appear with the edition.

Some examples of the information that might be attached to a DTS document follow:

- Source Information
- Author
- Title
- Date / Place of Original Creation
- Repository / Cataloging Info
- Publication History

- Comparanda / Related Documents
- Physical Description
- Language
- Provenance
- Mentioned People / Places / Events
- Surrogates (Editions, Images, Translations)
- Edition Information
- Editor
- Publication Info
- Language
- Source(s)
- Related Editions
- Revision History
- Attestations (may be edition-dependent)

The list is not comprehensive, but represents many of the kinds of *metadata* that can travel along with a TEI document. Much, if not all of it, might be represented in the document itself, but may also usefully be extracted and represented in a more generic form both in order to facilitate discovery in an API such as DTS, and to contextualize the document in a knowledge graph. Moreover, this list of metadata avoids making any strong assumption about the type of textual sources it can be applied to, and ideally aims to be applicable to any text.

If these types of metadata are reasonable candidates for representation in DTS, then we must consider how best to represent them. DTS allows for both basic Dublin Core and for an extended metadata set, with the idea that a client can choose whether to use only the *simple* properties from Dublin Core, or more specialized and sophisticated RDF data. Some possible mappings are listed in Table 1.

Examples

In order to exemplify the usage of the proposed document metadata scheme, let us consider now three examples (the first two speculative, the third one real).

1. A library has decided to publish OCR transcripts in TEI/XML format of its digitized collections as a DTS-compliant repository. These collections contain mostly editions and translations of Greek texts, including the OCR transcript of H. Boese's German translation of Proclus' *Institutio Physica*, entitled *Die mittelalterliche Übersetzung der Stoicheiosis phusike des Proclus*. In this case, source information describes Boese's translation, while edition information relates to the TEI/XML encoding of the OCR transcript.

The `dts:extensions` returned by the DTS document endpoint could provide contextual information about this text by using properties from the document metadata scheme described above:

Metadata	**RDF Representation**
Source Information	
Author	dc:creator, lawd:responsibleAgent
Title	dc:title
Date of Original Creation / Publication	dc:created, fabio:hasPublicationDate, fabio:hasPublicationYear
Place of Original Creation / Publication	lawd:origin, fabio:hasPublicationPlace
Repository / Cataloging Info	dc:identifier, crm:P1_is_identified_by
Publication History (the bibliography for a given primary source, e.g. papyrus or inscription)	dc:bibliographicCitation, list of lawd:Citation
Comparanda / Related Documents	dc:relation
Physical Description	dc:medium
Language	dc:language
Provenance (only applies to text-bearing artefacts)	lawd:foundAt, dc:provenance
Mentioned People / Places / Events	list of lawd:Attestation
Surrogates (Editions, Images, Translations)	dc:hasVersion
FRBR hierarchy of a text (Work, Expression, Manifestation, Item), where applicable	frbr:realization, frbr:embodiment, frbr:exemplar, fabio:isManifestationOf, fabio:isRealizationOf, fabio:isPortrayalOf
Edition Information	
Editor	dc:contributor, lawd:responsibleAgent
Publication Info	dc:bibliographicCitation
Source(s)	dc:source
Related Editions	dc:relation
Revision History	list of prov:Activity

Table 1. Metadata in the Document Metadata Scheme and their RDF representation.

```
"dts:extensions": {
 "@context": {
   "lawd": "http://lawd.info/ontology/",
   "frbr": "http://purl.org/vocab/frbr/core#",
   "fabio": "http://purl.org/spar/fabio",
   "ecrm": "http://erlangen-crm.org/current-version"
 },
 "ecrm:P1_is_identified_by": {
   "rdfs:label": "urn:cts:greekLit:tlg4036.tlg006.boese",
   "ecrm:P2_has_type": "http://purl.org/hucit/kb/types/CTS_URN"
 },
 @id: "https://library.org/texts/urn:cts:greekLit:tlg4036.tlg006.boese",
 @type: "fabio:DigitalItem",
 "fabio:isPortrayalOf": "http://purl.org/hucit/kb/works/3822",
 "frbr:exemplarOf": "https://library.org/editions/
                            urn:cts:greekLit:tlg4036.tlg006.boese",
  "dc:source": {
    @id: "https://library.org/editions/urn:cts:greekLit:tlg4036.tlg006.boese",
    @type: "lawd:Edition",
    "fabio:hasPublicationYear": "1958",
    "dc:medium": "printed edition",
    "fabio:hasPlaceOfPublication": "http://www.geonames.org/2950159/berlin",
    "dc:language": "de",
    "dc:title": "Die mittelalterliche Übersetzung der Stoicheiosis phusike des
         Proclus",
    "dc:bibliographicCitation": "Boese, H., Die mittelalterliche Übersetzung der
         Stoicheiosis phusike des Proclus (Deutsche Akademie der Wissenschaften
         zu Berlin, Institut für griechisch-römische Altertumskunde,
         Veröffentlichungen 6), Berlin: Akademie Verlag, 1958.",
    "fabio:isManifestationOf": "http://purl.org/hucit/kb/works/3822"
  }
}
```

2. Papyri.info is an aggregator of digital editions of papyri coming from various databases. Any given text may have connected metadata, translations, previous or following editions, and images. Disentangling all of these interrelated physical, print, and digital artefacts can be a real challenge. The Papyri.info system avoids trying to model the precise relationships of the various editions and metadata records that may pertain to a document, simply connecting them using the Dublin Core relation property. But, a more sophisticated representation of these relationships could be developed, perhaps using the CiTO ontology. The example below imagines some of the ways the metadata records for the papyrus P.Col 8 237 could be exposed in JSON-LD as DTS extended metadata:

```
"dts:extensions": {
 "@context": {
   "cito": "http://purl.org/spar/cito",
   "dc": "http://purl.org/dc/terms/",
   "ecrm": "http://erlangen-crm.org/current-version",
   "lawd": "http://lawd.info/ontology/",
 },
 @id: "https://papyri.info/ddbdp/p.col;8;237/source",
 @type: "lawd:Edition",
 "dc:source": {
  @id: "https://papyri.info/ddbdp/p.col;8;237/work", // URI representing the
                                                        print edition
  "cito:isCitedBy": "http://papyri.info/biblio/65200",
  "dc:source" {
   @id: "https://papyri.info/ddbdp/p.col;8;237/original", // URI representing
                                                             the primary source
   @type: "lawd:WrittenWork",
   "rdfs:seeAlso": "http://www.trismegistos.org/text/10553",
   "lawd:origin": "https://pleiades.stoa.org/places/737081",
   "lawd:foundAt": "https://pleiades.stoa.org/places/737081",
   "ecrm:P62i_is_depicted_by": "http://papyri.info/ddbdp/p.col;8;237/images",
   "dc:created": "3 June, 381 or 382"
  }
 }
}
```

3. Beta maṣāḥəft (BM) (Beta Maṣāḥəft: Manuscripts of Ethiopia and Eritrea n.d.) is a research environment for the manuscript tradition of Ethiopia and Eritrea. BM's data model makes a distinction between written artefacts (manuscripts, inscriptions, etc.) on one side, and textual and narrative units on the other.[3] The project exposes all available texts (editions, transcriptions, etc.) by means of a DTS API, and organizes them in a non-hierarchical way. Its DTS API makes use of `dts:extensions` to publish rich metadata about texts in BM in a structured and interoperable format (i.e., JSON-LD).

Inscriptions are also presented in this way, which helps to represent complex situations with minimal effort. For example, the edition of the Greek text of the pseudo-trilingual royal inscriptions (Bausi & Liuzzo 2018) *RIE 270*[4] has, as sources, two inscriptions:

```
"dc:source": [ {
  "fabio:isManifestationOf": "https://betamasaheft.eu/RIE185and270",
  "@type": "lawd:AssembledWork",
  "@id": "urn:dts:betmasMS:RIE185and270"
}, {
  "fabio:isManifestationOf": "https://betamasaheft.eu/RIE185bisand270bis",
  "@type": "lawd:AssembledWork",
  "@id": "urn:dts:betmasMS:RIE185bisand270bis"
} ]
```

Each of these written artefacts has a transcription and the second, RIE185bisand270bis includes a statement that it is a copy of the first.

```
"dts:extensions": {
  "crm:P1_is_identified_by": [ "104961", "RIÉ 270bis", "RIÉ 185bis and 270bis", "
      RIÉ 185bis" ],
  "saws:isDirectCopyOf": "https://betamasaheft.eu/RIE185and270"
}
```

A user could then take the text of the edition that uses only one of the texts from both copies on stone, or part of the transcription of one specific stone.

5 Conclusions and Further Work

In this paper, we addressed the problem of the absence, in the emerging LOD graph, of fully resolvable URIs for texts and their citable units. We sketched out a possible solution consisting of three main components: first, a registry service to keep track of available text repositories against which text URIs can be resolved; second, an identifier resolution service that can resolve a URI pointing to a section of a text or document and to one or more digital versions of that text or document; and third, a

[3] See Orlandi (2013) for Textual and Narrative Units. A further discussion can be found in Liuzzo (2019).

[4] The URI of Beta Masaheft LIT4851greekRoyal is https://betamasaheft.eu/api/dts/collections?id=urn:dts:betmas:LIT4851greekRoyal.

document metadata scheme to represent the relations between texts in the registry, as well as between these texts and related external resources. The next necessary step, before the implementation of these services can start, will be to gather feedback from the wider community. Such a discussion will help us ensure the generalizability of our solution (and especially the document metadata scheme) to potentially any text.

Acknowledgements

The authors wish to thank the *Pelagios Commons* project and the Andrew W. Mellon Foundation for funding the Linked Texts working group, from which this chapter originated. Thanks also to Pietro Maria Liuzzo for his feedback on earlier manuscript drafts, and to the DTS community, with whom we have been discussing many of the issues we address in this paper.

Bibliography

Almas, Bridget, Larry Lannom, and Robert Tupolo-Schneck, *CTS Handles Proposal*, 2018 <https://github.com/rpidproject/cts-handles/blob/master/proposal.md>

Bausi, Alessandro, and Pietro Liuzzo, 'Inscriptions from Ethiopia. Encoding Inscriptions in Beta Maṣāḥəft: From Practice to Discipline', in *Crossing Experiences in Digital Epigraphy* (De Gruyter, 2018), 84–92 <https://doi.org/10.1515/9783110607208-007>

'Beta Maṣāḥəft: Manuscripts of Ethiopia and Eritrea', n.d. <https://betamasaheft.eu/> 'Publications' <https://www.betamasaheft.uni-hamburg.de/publications.html>

Bodard, Gabriel, Hugh Cayless, Mark Depauw, Leif Isaksen, Faith Lawrence, and Sebastian Rahtz, 'Standards for Networking Ancient Person Data: Digital Approaches to Problems in Prosopographical Space', *Digital Classics Online*, 2017, 28–43 <https://doi.org/10.11588/dco.2017.0.37975>

Cayless, Hugh, 'LAWD. An Ontology for Linked Ancient World Data', 2016 <https://github.com/lawdi/LAWD>

'Center for Hellenic Studies: Online Publications', n.d. <https://chs.harvard.edu/CHS/article/display/1166.browse-online-publications>

Clérice, Thibault, 'Les Outils CapiTainS, l'édition Numérique et l'exploitation Des Textes', *Médiévales*, 73.73 (2017), 115–131 <https://doi.org/10.4000/medievales.8211>

Clérice, Thibault, Matthew Munson, and Bridget Almas, *Capitains/Capitains.Github.Io: 2.0.0*, version 2.0.0 (Zenodo, 2017) <https://doi.org/10.5281/zenodo.570516>

De Santis, Annamaria, and Irene Rossi, *Crossing Experiences in Digital Epigraphy, From Practice to Discipline* (Berlin, Boston: De Gruyter, 2019) <https://doi.org/10.1515/9783110607208>

'Distributed Text Services', n.d. <https://w3id.org/dts>

Elliott, Thomas, Sebastian Heath, and John Muccigrosso, 'Prologue and Introduction', *ISAW Papers*, 7.1 (2014) <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/elliott-heath-muccigrosso/>

Hedges, Mark, Anna Jordanous, K. Lawrence, Charlotte Roueche, and Charlotte Tupman, ‘Computer - Assisted Processing of Intertextuality in Ancient Languages’, *Journal of Data Mining and Digital Humanities*, 2016

Isaksen, Leif, Rainer Simon, and Elton T.E. Barker, ‘Social Semantic Annotation with Recogito 2.’, in *DH2017 Conference Abstracts*, 2017 <https://dh2017.adho.org/abstracts/570/570.pdf>

Isaksen, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares, ‘Pelagios and the Emerging Graph of Ancient World Data’, in *Proceedings of the 2014 ACM Conference on Web Science - WebSci ’14* (New York, New York, USA: ACM Press, 2014), 197–201 <https://doi.org/10.1145/2615569.2615693>

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych, ‘The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation’, in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (Association for Computational Linguistics, 2018), 5–9 <http://tubiblio.ulb.tu-darmstadt.de/106270/>

Liuzzo, Pietro Maria, *Digital Approaches to Ethiopian and Erithrean Studies* (Wiesbaden: Harrassowitz Verlag, 2019)

Middle, Sarah, ‘Identifying Research Methods for the Use and Production of Linked Ancient World Data’, 2018 <https://hcommons.org/deposits/item/hc:22031/>

Oberreither, Bernhard, ‘INTRO - an Intertextual Relationships Ontology for Literary Studies’, 2019 <https://github.com/BOberreither/INTRO>

Orlandi, Tito, ‘A Terminology for the Identification of Coptic Literary Documents’, *Journal of Coptic Studies*, 15 (2013), 87–94 <https://doi.org/10.2143/JCS.15.0.3005414rmin>

Peroni, Silvio, and David Shotton, ‘The SPAR Ontologies’, in *Artificial Intelligence and Soft Computing*, ed. by Leszek Rutkowski, Rafał Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz, and Jacek M. Zurada (Cham: Springer International Publishing, 2018), 119–36 <https://doi.org/10.1007/978-3-030-00668-6_8>

‘Perseus Catalog, Institutio Physica’, n.d. <https://catalog.perseus.org/catalog/urn:cts:greekLit:tlg4036.tlg006>

Rabinowitz, Adam, Ryan Shaw, Sarah Buchanan, Patrick Golden, and Eric Kansa, ‘Making Sense of the Ways We Make Sense of the Past: The Periodo Project’, *Bulletin of the Institute of Classical Studies*, 59.2 (2016), 42–55 <https://doi.org/10.1111/j.2041-5370.2016.12037.x>

Romanello, Matteo, and Michele Pasin, ‘Using Linked Open Data to Bootstrap a Knowledge Base of Classical Texts’, 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.*, ed. by Alessandro Adamou, Enrico Daga, and Leif Isaksen, CEUR Workshop Proceedings (CEUR-WS.org, 2017), MMXIV, 3–14 <http://ceur-ws.org/Vol-2014/paper-01.pdf>

‘SAWS The Ontology’, n.d. <http://www.ancientwisdoms.ac.uk/method/ontology/>

Smith, Neel, ‘Citation in Classical Studies’, *Digital Humanities Quarterly*, 3.1 (2009) <http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html>

‘The Canonical Text Service (CTS) : The CITE Architecture’, 2019 <http://cite-architecture.org/cts/>

‘Trismegistos : An Interdisciplinary Portal of the Ancient World’, n.d. <https://www.trismegistos.org/>

Wagner, Andreas, ‘The School of Salamanca - The Web Application, API Documentation’, 2019 <https://github.com/digicademy/svsal/blob/master/docs/API.md>

Witt, Jeffrey C., ‘Digital Scholarly Editions and API Consuming Applications’, in *Digital Scholarly Editions as Interfaces*, ed. by Roman Bleier, Martina Bürgemeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider (Norderstedt: Books on Demand, 2018), 219–247

SPEEDy. A Practical Editor for Texts Annotated with Standoff Properties

Iian Neill, Desmond Schmidt

Abstract

Standoff properties can be used to record textual properties or annotations that may freely overlap and need not conform to a context-free grammar. In this way they avoid the 'overlapping hierarchies' problem inherent in markup languages like XML. Instead of embedding markup tags directly into the text stream, standoff properties are stored separately, and refer to positions in the text where each property starts and ends. However, this has the effect of tightly binding the properties to the text, and hence any change in the underlying text invalidates them. This limitation usually makes this method impractical in cases where the text is mutable, and is mostly used when the text is already fixed or proofread to a high standard. However, if it did become feasible to use standoff properties on mutable texts, this method could also be used in the process of text production, on dynamically evolving texts, such as emails, forum messages, personal notes and even drafts of academic papers. Digitised transcriptions of historical documents, whether produced manually or through OCR, could then be easily corrected at an earlier stage of typographic correctness. By overcoming the overlapping hierarchies problem this technique thus offers the prospect of significant productivity gains for producing digital editions, as well as a new mode of engagement for annotation. This paper describes the SPEEDy editor, a practical realisation of this technique. It outlines the editor's foundational concepts, its standoff properties model, and its main interface features.

1 Introduction

SPEEDy (Neill 2020a) was created as part of a personal project to transcribe and annotate the letters of Michelangelo (1474–1564). As this project progressed, it became apparent that managing the numerous references to the people, places, and events mentioned in the letters required not only an annotation system, but also an editor, to allow for the updating of the content, its formatting, and annotations, in one place.

Annotations are inherently overlapping. Ranges of text may be arbitrarily selected by the user and annotations applied to them. However, the underlying HTML markup of a web page *is* strictly nested and forbids overlap. This creates a potential conflict

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 45–54.

between the two systems of markup. One is internal, describing layout and text structure, the other external or out-of-line, which points to it. Standard systems for specifying external annotations, such as the Open Annotation Model and annotation clients based on it, Recogito (n.d.) and Hypothesis.is (2011-), are limited in that they cannot specify arbitrary ranges within HTML (Sanderson et al., 2013, 2.1.4). They also assume that the text being annotated is immutable, since any change to it would potentially invalidate the annotations that refer to it (Tennison 2012, 7.2). Internal annotation systems such as RDFa (Herman et al. 2015) and Microformats (2020) are not subject to this limitation; however, they can only provide attributes for correctly nested HTML elements, and are hence unable to represent overlapping structures.

However, the distinction between the markup of the text being annotated and that of the annotations is artificial, since even formatting markup is partly semantic, and is thus a kind of annotation. For example, a paragraph element in HTML is not merely a formatting construct (e.g., left-aligned or with double line-spacing etc.) but also indicates that the enclosed text is a separate sequence of thoughts, distinct from the paragraphs that precede or follow it.

It has long been known that the transcription of historical sources often requires the description of overlapping structures (Renear et al., 1993). In markup languages like SGML and XML a context-free grammar is defined to regulate the syntax of the tags and text, the so-called DTD or schema. But this requirement, perfectly suited to the composition of new business documents, leads to problems when transcribing historical texts. For example, in holographs, revisions such as the joining of two paragraphs, overlapping underlining, and the conflicting demands of semantic, formatting and sentence structures, suggest that an approach seeking to redefine all markup as a kind of annotation might solve these conflicts in a new way. Buzzetti (2002) argued some time ago that *weakly embedded markup*, by which he seems to mean a kind of standoff annotation describing possibly overlapping ranges in the underlying text, may provide a better representation of textual structure and meaning than strongly embedded markup languages like XML. However, to be practicable, this would mean that the underlying text of the annotations would have to be editable and fluid, not fixed and immutable, and that the pointers into the text by the annotations, like the text itself, would have to be dynamically updateable.

SPEEDy differs from standoff markup editors like MUP (Glass & Di Eugenio 2002), the MATE annotation workbench (McKelvie et al., 2001), or the repository approach to standoff markup used in *Knora* (Rosenthaler et al., 2019), all of which preserve the hierarchical structure of the original embedded markup, and require an immutable base text. It also differs from web-based tools such as T-PEN (n.d.) for manuscript transcription, or editors like Quill (n.d.), which focus on editing, rather than annotation. SPEEDy instead combines features of both web-based editors and annotation clients in a single design.

The rest of this paper is divided into five sections. The first discusses the model for representing text in the editor, the second describes the data model for annotations, and the third describes the key features of the user interface. The fourth section describes future plans for the editor and the fifth draws some conclusions for its use in digital humanities projects.

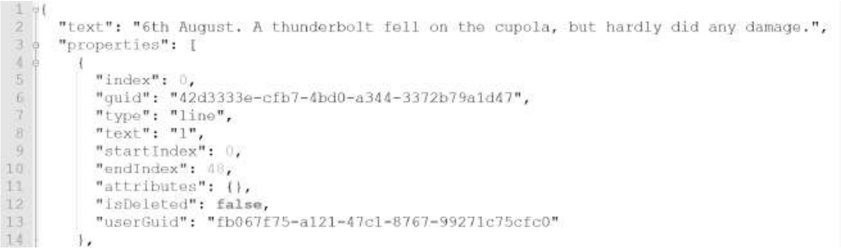
2 Representing Text

SPEEDy is built on the notion of *cells,* which are individual characters arranged in a linked list, which forms the text stream. The concept of the text stream itself is expressed in two modes: *in-stream* and *out-of-stream* cells. An *in-stream cell* contains a single character and the sequence of *in-stream* cells collectively forms the plain text output. An *out-of-stream* cell is content which is visible in the editor panel, but which is not included in the plain text output. The assignment of text to one of these two modes is a matter of choice by the human editor. For example, a page number may or may not be considered part of the plain text output. Other examples of *out-of-stream* text include footnote labels, end-of-line hyphenation, and Leiden Convention symbols.

Since SPEEDy is a web-based editor, it uses the `<span>` element of HTML to implement individual cells, one character per cell. Any properties that the text acquires when loaded or during editing are stored in an array of standoff property JavaScript objects, which are linked to the `<span>` elements at the start and end of the annotation range. Therefore, annotations can move about naturally as the text is edited. Blocks of text such as paragraphs and other textual divisions are represented using the `<div>` element. No other HTML elements are used, or needed. Originally SPEEDy only supported `<spans>s`, linked together to form the text content. This kept the editor simple, and allowed the easy application of overlapping annotations. However, the need for paragraph level formatting such as left, right, centre, or justified text alignment, and indentation required the segmentation of this simple linked-list model of text into blocks. Navigation between blocks is now performed not by directly navigating the linked list structure, but via a tree-walking algorithm, which traverses the sequence of text cells as if it were contiguous. In other words, every index in the text refers to a unique location, either between two cells, or at the start or end of the entire text.

This distinction between blocks and characters is based on the assumption that, while semantic annotations are not inherently hierarchical, annotations describing layout require some kind of shallow hierarchy for representation in real-world formatting systems like HTML.

In the database the *in-stream* text is stored as plain text, with no markup of any kind. The *out-of-stream* text, and the markup which was contained in the attributes



```
{
  "text": "6th August. A thunderbolt fell on the cupola, but hardly did any damage.",
  "properties": [
    {
      "index": 0,
      "guid": "42d3333e-cfb7-4bd0-a344-3372b79a1d47",
      "type": "line",
      "text": "1",
      "startIndex": 0,
      "endIndex": 48,
      "attributes": {},
      "isDeleted": false,
      "userGuid": "fb067f75-a121-47c1-8767-99271c75cfc0"
    },
```

Figure 1. JSON export of SPEEDY annotation.

of the spans and blocks during editing, is stored as a set of properties, which point to the text in accordance with the standoff properties model. When a document is called up for editing again, the text and properties are recombined to generate the blocks and spans, so completing the cycle.

3 Representing Annotations

As described above, SPEEDy represents both formatting markup and semantic annotations as standoff properties that may freely overlap (Schmidt 2016). The more familiar term standoff markup refers to markup tags that represent a strict hierarchical text structure, conforming to a context-free grammar, which have been removed from the text and stored separately. Standoff properties, on the other hand, do not conform to a grammar, and do not have an intrinsic hierarchical structure. In XML, the data model for tags consists of elements and their attributes. Likewise, in SPEEDy, the standoff properties have their own data structure, which is described in this section.

One way to see the structure of annotations is to export a SPEEDy document to a JSON format. This shows how the document is represented as a plain text string, and its properties as an array. Figure 1 shows an example of this output, showing a short text and one of its standoff properties.

The essential fields of each property are:

- `type`. This represents the name of the annotation. In XML this would be the element name.
- `startIndex`. An integer representing the index position of the first character of the text range: 0 ≦i < n, where *i* is the index and *n* is the length of the text.
- `endIndex`. An integer representing the index position of the last character of the text range: `startIndex` ≦i < n, where *i* is the index and *n* is the length.
- Not all annotations will describe a simple range of text. Two other possibilities are annotations that refer to a specific point in the text, and those that refer to the document as a whole.

- A null value for endIndex signifies that the property represents a point between two indexes, called here a *zero-width annotation* (ZWA).
- A null startIndex signifies that the property is metadata associated with the text as a whole, and does not refer to any specific position or range within it.

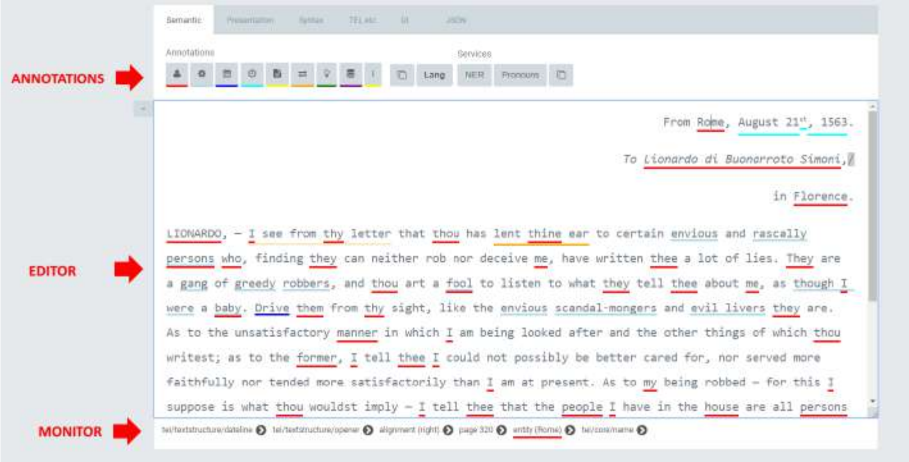
The other main fields of an annotation are:

- value. Any value which can be assigned to an annotation, depending on its function. For example, an entity annotation (i.e., for some kind of name) would use value to store the entity's identifier, while a font annotation would store the font name. There is no native *entity* or *font* annotation in SPEEDy, since value refers to something that is externally defined.
- attributes. A schema-less array field that is functionally equivalent to the attributes of an XML element. The choice between value and attributes for storing annotation data is one of convenience: typically, if it only makes sense for an annotation to have a single value then it is better to use value; if multiple values are required, then attributes should be preferred. An example of the latter is a Part of Speech annotation, which might contain data like gender, case, person, etc. Attributes are also used in LMNL, another markup formalism that uses standoff properties (Piez, 2015).
- text. For regular annotations of text ranges, the text field contains a copy of the text in its range in the text proper (arbitrarily truncated at 100 characters by default, but overridable). This facilitates reading the data either in the JSON output or in the data store. But the true object of the annotation is always the sequence of characters in the text starting at startIndex and ending at endIndex. For zero-width annotations, this field is used to store the *out of stream* text that denotes the inter-index position of the ZWA. For example, if a footnote were modelled as a ZWA, the footnote number or label would be stored in the text field.
- guid. The standoff property's unique identifier. No format is specified, but supplying an identifier for each annotation facilitates storage and querying in the external annotation store.

The index, isDeleted and userGuid fields are used internally by SPEEDy, and need not be explained here.

4 The Editor User Interface

The user interface is important because it is the window through which the user sees and manipulates the underlying textual and annotation data models. As explained above, text formatting and semantic annotations are merged in SPEEDy, and this



From Rome, August 21st, 1563.

To Lionardo di Buonarroto Simoni,

in Florence.

LIONARDO, – I see from thy letter that thou has lent thine ear to certain envious and rascally persons who, finding they can neither rob nor deceive me, have written thee a lot of lies. They are a gang of greedy robbers, and thou art a fool to listen to what they tell thee about me, as though I were a baby. Drive them from thy sight, like the envious scandal-mongers and evil livers they are. As to the unsatisfactory manner in which I am being looked after and the other things of which thou writest; as to the former, I tell thee I could not possibly be better cared for, nor served more faithfully nor tended more satisfactorily than I am at present. As to my being robbed – for this I suppose is what thou wouldst imply – I tell thee that the people I have in the house are all persons

Figure 2. Screenshot of SPEEDy as integrated into Codex.

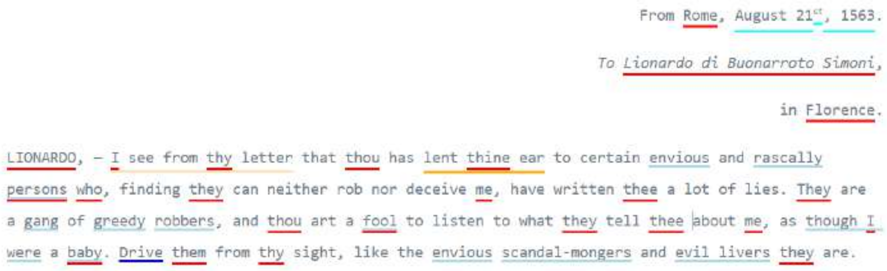
increase in cognitive load must be carefully handled so as not to overwhelm the user with too much detail (Nielsen 1993, 129f).

The annotations available in SPEEDy will vary greatly from one project to another, and are therefore fully configurable via a JSON file. This can include other vocabularies such as TEI (Neill 2020a). Although SPEEDy is an independent editor, it can be integrated in a variety of applications. The following screenshots are taken from the integration of SPEEDy in the Codex environment (Neill & Kuczera 2019), which provides visualisations of annotations across a document collection. It is based on a graph meta-model stored in Neo4j. However, all the visualisations provided below refer only to SPEEDy and do not concern the Codex system.

The editor interface is divided into three main sections: the annotation toolbar, the editor window and the monitor toolbar, as shown in Figure 2 (Carden 1913, 320; Neill 2020b).

Figure 3 shows a close-up of the annotated text of the above Michelangelo letter (Carden 1913, 320). Semantic annotations are visualised with coloured lines which are positioned below the line so that overlaps between properties are obvious. Also it is clear that the annotations can coexist perfectly with textual properties such as the right aligned text of the opening section of the letter.

Figure 4 shows the monitor toolbar, which changes whenever the user clicks on an annotation in the editor window. It displays any annotations that enclose the current cursor position. In Figure 2 the cursor is on the word Rome, which is a named entity that is itself inside several other annotation ranges. Hovering over an annotation in



From Rome, August 21st, 1563.

To Lionardo di Buonarroto Simoni,

in Florence.

LIONARDO, – I see from thy letter that thou has lent thine ear to certain envious and rascally persons who, finding they can neither rob nor deceive me, have written thee a lot of lies. They are a gang of greedy robbers, and thou art a fool to listen to what they tell thee about me, as though I were a baby. Drive them from thy sight, like the envious scandal-mongers and evil livers they are.

Figure 3. Detail of a text in SPEEDy showing the use of coloured underlines for visualising overlapping semantic annotations.

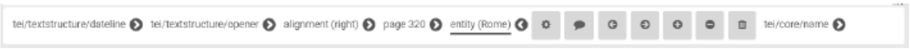


Figure 4. The monitor bar.

the monitor toolbar highlights both the annotation name and its range in the text. To the right of each annotation is a right-facing chevron which, when clicked, displays a list of buttons for editing the annotation (Neill 2020b). Figure 5 shows this in more detail.

Counting from left to right the buttons perform the following functions: [1] launches a window for changing the entity reference; [2] enables a comment to be added to the annotation (the comment itself is a standoff property document); [3] shifts the annotation to the left by one character; [4] shifts the annotation to the right by one character; [5] expands the annotation by one character; [6] retracts the annotation by one character; and [7] deletes the annotation. Although these interactions are fairly complex the buttons only appear when the right chevron is clicked, and also only apply to the currently selected annotation.

Figure 6 (Landucci & del Badia 1927, 299; Neill 2020b) shows two examples of out-of-stream text. Recalling the discussion in section 1 above, out-of-stream text is visible in the editor but is not included in the plain text output. The first example is a zero-width hyphen annotation which appears in the middle of the word *paternoster.*

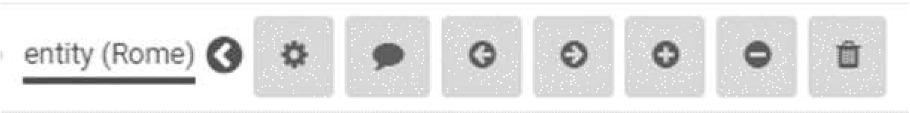
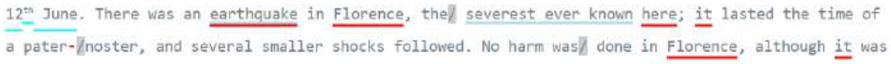
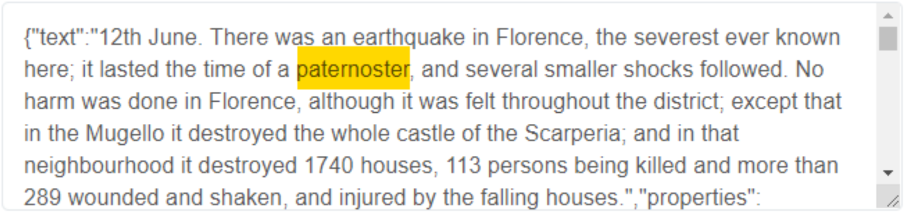


Figure 5. Controls in the monitor for manipulating an annotation.



12th June. There was an earthquake in Florence, the/ severest ever known here; it lasted the time of
a pater-/noster, and several smaller shocks followed. No harm was/ done in Florence, although it was

Figure 6. Examples of *out-of-stream* text.



{"text":"12th June. There was an earthquake in Florence, the severest ever known here; it lasted the time of a paternoster, and several smaller shocks followed. No harm was done in Florence, although it was felt throughout the district; except that in the Mugello it destroyed the whole castle of the Scarperia; and in that neighbourhood it destroyed 1740 houses, 113 persons being killed and more than 289 wounded and shaken, and injured by the falling houses.","properties":

Figure 7. Example showing that the zero-width soft hyphen annotation in 'pater-noster' is not rendered in the output text stream.

The second is the line-break annotation, which is shown as a greyed-out forward slash. A hyphen annotation is used to represent a soft hyphen, which appears in the source but has only typographical, not linguistic, significance (which would be a hard hyphen). Including soft hyphens in the text content complicates the process of exact text matching, syntax analysis and lemmatisation. By modelling the hyphen annotation as a zero-width annotation, it can be displayed at its anchor location as an out-of-stream red-coloured hyphen.

Figure 7 (Landucci & del Badia 1927, 299; Neill 2020b) shows that the word *pater-noster* is not hyphenated in the recorded plain text section of the JSON output for this document. The same technique of using out-of-stream text is also applied to the line-break annotations.

5 Future Plans

Further developments planned for SPEEDy include:

- the ability to import and export TEI-XML text;
- support for scripts with contextual letterforms, such as Arabic and Syriac;
- support for Unicode planes other than the basic multilingual plane;
- the virtualisation of the editor window memory space to handle texts of any size;
- additional monitor sections for linking the editor to real-time visualisations;
- examination of the practicality of annotating both plain text and XML structures; and
- handling of multi-version and multi-layered texts, such as edited holographs.

6 Conclusion

SPEEDy has demonstrated the practicality of an editor based on standoff properties. Offering the user the possibility of editing texts that can be enriched with a customisable set of textual properties and annotations, and yet saved in a form that cleanly separates them from the base text, provides greater flexibility in the tasks of transcription and annotation. Corrections to OCR-scanned editions or transcribed original sources can now include markup and annotation from the start, rather than be introduced at a later stage in the production process.

Bibliography

Anon, ‘Knora Documentation’, *Knora* <https://docs.knora.org/>

Buzzetti, Dino, ‘Digital Representation and the Text Model’, *New Literary History*, 33 (2002), 61–88

Carden, Robert W., *Michelangelo Buonarroti: Michelangelo: A Record of His Life as Told in His Own Letters and Papers* (London: Constable & Co, 1913)

Glass, Michael, and Barbara Di Eugenio, ‘MUP: The UIC Standoff Markup Tool’, in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, SIGDIAL (Association for Computational Linguistics, 2002), 37–41 <https://doi.org/10.3115/1118121.1118126>

Herman, Ivan, Ben Adida, Manu Sporny, and Mark Birbeck, ‘RDFa 1.1 Primer - Third Edition Rich Structured Data Markup for Web Documents’, 2015 <https://www.w3.org/TR/xhtml-rdfa-primer/>

Hypothesis, founded by Dan Whaley, [2011-] <https://web.hypothes.is/>

Landucci, L, and I del Badia, *A Florentine Diary from 1450 to 1516. Trans. by Alice de Rosen Jervis* (New York: J.M. Dent, 1927)

McKelvie, David, Amy Isard, Andreas Mengel, and Morten Baum Møller, ‘The MATE Workbench –an Annotation Tool for XML Coded Speech Corpora’, *Speech Communication*, 33.1–2 (2001), 97–112

‘Microformats’, 2020 <http://microformats.org/>

Neill, Iian, ‘Editor Demo’, 2020a <https://argimenes.github.io/standoff-properties-editor/>

———, ‘Standoff-Properties-Editor’, 2020b <https://github.com/argimenes/standoff-properties-editor>

Neill, Iian, and Andreas Kuczera, ‘The Codex – an Atlas of Relations’, *ZfdG - Zeitschrift Für Digitale Geisteswissenschaften*, Sonderband 4 (2019) <http://www.zfdg.de/sb004_008>

Nielsen, Jakob, *Usability Engineering* (Boston: Academic Press, 1993)

Piez, Wendell, ‘TEI in LMNL: Implications for Modeling’, *Journal of the Text Encoding Initiative*, 8 (2015) <https://journals.openedition.org/jtei/1337>

‘Quill’, *Quill* <https://quilljs.com/>

Renear, Alan, Elli Mylonas, and David Durand, ‘Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies’, 1993 <http://cds.library.brown.edu/resources/stg/monographs/ohco.html>

Rosenthaler, L., B. Geer, and T. Schweizer, ‘Knowledge Organization, Representation, and Annotation’, *Knora*, 2019 <https://www.knora.org/>

Sanderson, Robert, Paolo Ciccarese, and Herbert Van de Sompel, ‘Open Annotation Data Model’, *Open Annotation Data Model*, 2013 <http://www.openannotation.org/spec/core/>

Schmidt, Desmond, ‘Using Standoff Properties for Marking-up Historical Documents in the Humanities’, *It – Information Technology*, 58.H.2 (2016), 63–59

Recogito, developed by Elton Barker, Leif Isaksen, Rebecca Kahn, Rainer Simon, and Valeria Vitale, n.d. <https://github.com/pelagios/recogito2>

Tennison, Jeni, ‘Best Practices for Fragment Identifiers and Media Type Definitions’, 2012 <https://www.w3.org/TR/fragid-best-practices/>

‘T-PEN Version 2.8’, *T-PEN* <http://www.t-pen.org/TPEN/>

The Power of OCHRE's Highly Atomic Graph Database Model for the Creation and Curation of Digital Text Editions

Miller C. Prosser, Sandra R. Schloen

Abstract

The Online Cultural and Historical Research Environment (OCHRE) is a research database platform that provides a suite of tools to aid in the curation of digital text editions. The power and flexibility of the OCHRE system is predicated on the underlying data model, which is constructed as a graph database. OCHRE is based on a semi-structured, item-based data model where data are atomized into granular items and organized through hierarchical arrangement and cross-cutting links. In this paper, we describe OCHRE's graph data model and demonstrate how this approach revolutionizes digital philology.

1 Introduction

The Online Cultural and Historical Research Environment (OCHRE) is a research database platform that provides a suite of tools to aid in the curation of digital text editions.[1] But that was not the original mandate of the program. OCHRE was created as a tool for archaeological data management.[2] Because no two archaeologists can be compelled to agree on a common system for excavation or on a single controlled vocabulary for describing their data, OCHRE was created to be customizable. Further, because an archaeologist typically needs to describe a wide variety of data, from a single botanical sample to an entire watershed region, OCHRE was designed to manage data at any level of abstraction or observation – indeed, at multiple levels of abstraction, or with multiple observations. Finally, because archaeology creates data of many types – such as images, geo-spatial (GIS), and daily journal entries to

[1] We use the term curation to refer to the activities and processes performed to capture, create, edit, publish and archive data—essentially the entire data life-cycle.

[2] Sandra R. Schloen and J. David Schloen invented OCHRE as a data management system for David Schloen's archaeological research. As a trained software developer, Sandra Schloen implemented their plan as a database platform. Over decades of use and decades of technological advancement, OCHRE has evolved through many phases. For more information on OCHRE in general, see Schloen and Schloen (2012).

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 55–71.

name only a few – OCHRE was built to integrate any and all types of project data.[3] The database, the underlying data model, and the user interface evolved to become highly flexible, generic, and extensible. For these reasons, it became clear that the system could be applied to a variety of other research domains. Based as it is at the Oriental Institute (OI) of the University of Chicago, the next logical application of the OCHRE system was the field of philology, one of the other core research areas at the OI. The same affordances granted the archaeologist are available to the philologist. Researchers are free to use a set of descriptive terms that are recognized in their specific area of study or to create their own knowledge representation vocabulary. Observations can be made by multiple authors or editors at any level of detail, from a single grapheme to an entire corpus. Data of all types are integrated in a common platform, allowing the presentation of text images, bibliography, and commentary along with textual data.

The power and flexibility of the OCHRE system is predicated on the underlying data model, which is constructed as a graph database. OCHRE is based on a semi-structured, item-based data model, where data are atomized into granular items and organized through hierarchical arrangement and cross-cutting links. In this paper, we describe OCHRE's graph data model and demonstrate how this approach revolutionizes digital philology.

2 The OCHRE Data Model

As mentioned already, data in OCHRE is highly atomized. By atomized, we mean that data is broken down into minimal meaningful parts, each of which is stored as a separate XML document.[4] Data is not stored in a tabular format – the data model used by a typical relational database. Data is not stored in fully composed and marked-up documents – the data model often used as the default approach for textual research. Instead, in OCHRE, items are arranged in a semi-structured, hierarchical model, an arrangement that is supplemented by cross-cutting links between items. The result is an integrated graph of data nodes. These nodes are categorized according to a generic upper ontology that specifies the classes and relationships between the nodes. The high-level data categories in OCHRE are the following: Agents, Bibliography, Concepts, Dictionary units, Periods, Resources, Spatial units, Texts, and Writing systems. Agents are people, real or fictional, ancient or modern, including even

[3] On issues related to using digital tools for archaeological data management, see Prosser (2020), "Digging for Data."

[4] OCHRE is implemented using the Tamino XML database, the first enterprise-level native-XML database (developed by Software AG, Germany). While data in OCHRE is stored as documents, the database is more accurately described as a semi-structured graph database rather than as a document-oriented database because it is characterized by relationships between highly atomized database items.

project team members. Resources are images, PDFs, audio files, or any other external file. Spatial units are any items, real or otherwise, that can be contextualized in space, meaning that they can be organized according to their location. A Spatial unit may have a latitude-longitude coordinate; it may be a single point or an entire region. Along with space, time is also data, which is recorded in OCHRE's Periods category. Even the controlled vocabulary of the project is stored as data. A hierarchy of variables and values that define the descriptive properties of all the other items forms a project Taxonomy. OCHRE manages a master taxonomy, from which projects may borrow to create their own local taxonomy. However, project personnel can customize their taxonomy to include unique variables and values needed to describe their research. Any OCHRE item can be described with properties as allowed by the project taxonomy. In addition, any OCHRE item can be linked to any other OCHRE item(s) using a variety of mechanisms. As we go on to define the details of the Text category of items below, keep in mind that, at its core, all data in OCHRE is a network of data organized within these high-level categories, or node classes, described by properties, and related by links.

2.1 Textual Data

Contrasting with what can be called the document model, wherein a series of string characters are stored in a sequence that corresponds to the layout of the text on the page, OCHRE's data model breaks down textual data into items that correspond to either words or graphemes.[5] Each item is uniquely identified and stored as a separate XML document. These items can be combined and organized to produce any variety of derived formats appropriate for viewing, analysis, publication, or even good old-fashioned printing. Each item can be addressed individually by the researcher: identified, commented upon, or reused in a variety of overlapping hierarchical contexts.

To make this point from another perspective, when guidelines from the Text Encoding Initiative (TEI consortium 2019) are used to record and describe the structure of a text, the result is a richly marked up single XML document for the entire text.[6]

[5] We use the linguistic term grapheme to refer to any minimal and meaningful unit of writing, more colloquially referred to as a letter, sign, accent, or punctuation. On the difference between a document model and a database model, see Schloen and Schloen (2014).

[6] We bring TEI into the discussion here because it is so commonly implemented in textual studies and because we want to emphasize that our use of XML is not the same as TEI-XML. In the simplest implementation of TEI markup, an entire text is stored as a single XML file. There is good reason for doing this and we have no criticism of this approach. It has utility in certain contexts. Textual data from OCHRE can be exported, then styled with an XSLT stylesheet to create well-formed TEI-XML. We allow the researcher to make this transformation to TEI to avoid imposing any single implementation of TEI on a researcher.

By contrast, in the OCHRE data model, we create a separate XML document for every word or grapheme or *minimal meaningful part* of the text. So, instead of one XML document per text, we work with hundreds or thousands of XML documents for each text.[7] The structure of the text is represented, in part, by hierarchical arrangement of these items. The hierarchy, itself, is a separate XML document that organizes its content. That is, the hierarchy has links to those items that it contains, and the items in turn link back to the (potentially many different) hierarchies in which they are contained. This approach allows any item to participate in multiple hierarchies. OCHRE represents this complex network of data in a natural and intuitive user interface. To the end user, a text view looks very much like a sequence of string characters.[8]

To illustrate the highly atomic and granular data model implemented in OCHRE, let us turn to the manner in which textual data is represented. First, we separate the idea of the object and the text. The object is what the paleographer refers to as the writing support, i.e., the surface or object on which the text is present. The object, a Spatial unit in OCHRE's ontology, has its own set of metadata properties and descriptions. Objects may have coordinates that represent where they were discovered – important for archaeology projects; or a designation to indicate where they are stored – useful for managing an inventory of objects in a museum context.

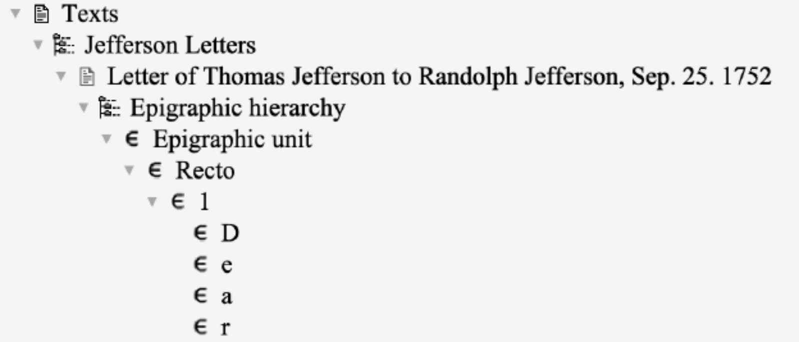
As distinct from the object, the Text item in OCHRE is composed of two major classes of items.[9] Specifically, a text consists of collections of epigraphic units and discourse units. The class of items called epigraphic units are organized hierarchically as recursive elements to describe the epigraphic layout of the text on the writing support, whether it be a folio, page, or ancient clay tablet. From broadest to most narrow, the recursive hierarchy of epigraphic units may be, as one example, Recto, Line 01, Latin uppercase D.

In addition to the epigraphic hierarchy, a text is defined using a discourse hierarchy, which represents a scholarly interpretation of the text. Any text may consist of multiple discourse hierarchies: for example, one for a word-by-word interpretation, one for a poetic analysis, or one for a syntactic analysis. A text may have multiple

[7] We admit that it may well be possible to implement a highly atomized and granular approach to texts using TEI. It may be possible to utilize the pointing mechanism of TEI to define a text as being composed of thousands of external files. However, in this scenario, one wonders what value is gained by encoding the XML files as TEI. The highly atomized approach we use in OCHRE is better implemented as a graph database. On the use of pointers in TEI see TEI Consortium 2019, ch. 16. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SAXP, last accessed April 2, 2019.

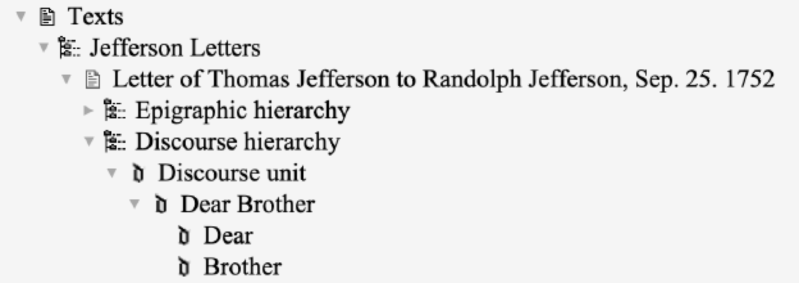
[8] OCHRE has a sophisticated import tool built in to atomize a text document of the ordinary kind into a network of highly granular but related parts. The user need only set a few options and click the *Import* button.

[9] Here we use the word "class" to refer to a component of an ontology, a category of similar concepts.



- Texts
  - Jefferson Letters
    - Letter of Thomas Jefferson to Randolph Jefferson, Sep. 25. 1752
      - Epigraphic hierarchy
        - Epigraphic unit
          - Recto
            - 1
              - D
              - e
              - a
              - r

Figure 1. Partial view of the epigraphic structure of a Jefferson Letter.



- Texts
  - Jefferson Letters
    - Letter of Thomas Jefferson to Randolph Jefferson, Sep. 25. 1752
      - Epigraphic hierarchy
      - Discourse hierarchy
        - Discourse unit
          - Dear Brother
            - Dear
            - Brother

Figure 2. Partial view of the discourse hierarchy of the Jefferson Letter.

discourse hierarchies that represent interpretations of various scholars.[10] A discourse hierarchy consists of discourse units such as paragraph, sentence, stich, phrase, clause, or word. The smallest discourse unit is usually the word. Any given discourse unit can be reused in multiple discourse hierarchies. The same XML file that represents the word, for example, may point to a discourse hierarchy representing poetic analysis and a discourse hierarchy representing syntactic analysis. Reuse of data in this fashion solves the problem of multiple overlapping hierarchies in textual data. Further, the database item that represents the word is linked to the epigraphic units that represent the graphemes that are its constituent parts. Stated inversely, a series of epigraphic units is linked to a discourse unit.

To use the semantics of graph theory, a text is an atomized network of nodes and edges. An epigraphic unit – whether it be a grapheme or one of the larger epigraphic

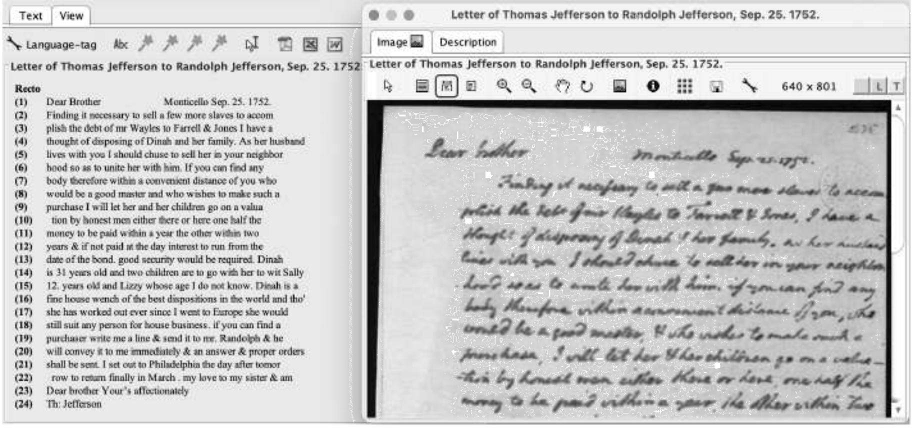
[10] Especially in the world of ancient texts, there are often disagreements about the interpretations of graphemes and words.

sections – is available for reuse and sharing among multiple texts, because it is simply a node in a network. These multiple contexts may be editions prepared by different editors, or even separate texts copied by ancient scribes. We demonstrate below a practical application of this approach. An epigraphic unit records these network connections as a list of links that represents the edges that point to other nodes to which it is related. Each epigraphic unit is a node that may have an unlimited number of edges, i.e., the pointers that record where in the text a given node exists. In other words, there is no proscribed definitional boundary to the set of nodes that can link to each other. The node that represents a given word may link to any number of other larger discourse units such as couplets, lines, sentences, or paragraphs. Further, that same word may link to competing scholarly interpretations of couplets, lines, sentences, or paragraphs. This highly atomized graph approach to textual data provides an elegant solution to the problem of representing the same data in overlapping or competing hierarchies.[11]

Although the underlying OCHRE structure is modelled as a graph, OCHRE uses hierarchical structures to organize the nodes (items) and edges (links), rather than using node-edge style visualizations commonly associated with network analysis and graph databases. (See the two figures above for sample hierarchies.) It is worth noting that there is no need to choose between a hierarchy and a graph. A hierarchy *is* a graph, and OCHRE exploits the advantages of graphs in general, and hierarchies more specifically.[12] The hierarchical arrangement of graphemes within words, within phrases/clauses, within sentences/lines, within pages, and so on, is an intuitive construct, and closely parallels how scholars naturally work with, and think about, their textual data (and lexical, taxonomic, archaeological, and many other types of research data). OCHRE also recombines, on demand, the highly atomic items into composite views that are displayed to the scholar in familiar formats.

[11] The so-called *problem* of multiple, overlapping hierarchies when representing texts is well attested in the literature, usually in the context of applying markup (e.g., TEI) to documents. For a discussion of this issue see Schloen and Schloen (2014). But multiple overlapping hierarchies are unproblematic in the context of representing texts using an item-based approach, where each hierarchy is treated as a discrete item and even, potentially, representative of a work of scholarship. That is, each epigraphic or discourse hierarchy is simply a node of a graph that links to many other nodes (epigraphic or discourse units) in conformance to carefully applied rules.

[12] According to Robinson et al (2015, 109) “A graph database’s structured yet schema-free data model” makes them “ideally applied to the modeling, storing, and querying of hierarchies…”. Examples are given of how to represent “cross domain models” (*ibid.*, 41ff), which is analogous to how OCHRE manages multiple overlapping hierarchies without difficulty using a graph approach.



Letter of Thomas Jefferson to Randolph Jefferson, Sep. 25. 1752

Recto
(1) Dear Brother Monticello Sep. 25. 1752.
(2) Finding it necessary to sell a few more slaves to accom
(3) plish the debt of mr Wayles to Farrell & Jones I have a
(4) thought of disposing of Dinah and her family. As her husband
(5) lives with you I should chuse to sell her in your neighbor
(6) hood so as to unite her with him. If you can find any
(7) body therefore within a convenient distance of you who
(8) would be a good master and who wishes to make such a
(9) purchase I will let her and her children go on a valua
(10) tion by honest men either there or here one half the
(11) money to be paid within a year the other within two
(12) years & if not paid at the day interest to run from the
(13) date of the bond. good security would be required. Dinah
(14) is 31 years old and two children are to go with her to wit Sally
(15) 12. years old and Lizzy whose age I do not know. Dinah is a
(16) fine house wench of the best dispositions in the world and tho'
(17) she has worked out ever since I went to Europe she would
(18) still suit any person for house business. if you can find a
(19) purchaser write me a line & send it to mr. Randolph & he
(20) will convey it to me immediately & an answer & proper orders
(21) shall be sent. I set out to Philadelphia the day after tomor
(22) row to return finally in March . my love to my sister & am
(23) Dear brother Your's affectionately
(24) Th: Jefferson

Letter of Thomas Jefferson to Randolph Jefferson, Sep. 25. 1752.

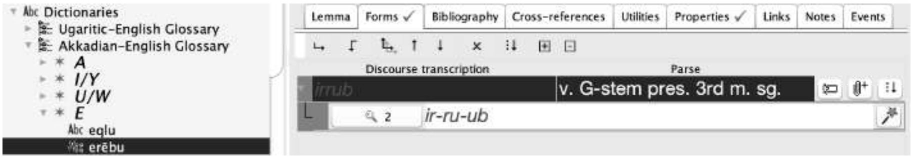
Figure 3. Recomposed document view of the Jefferson Letter example with associated image. (http://www.loc.gov/exhibits/jefferson/images/vc109a.jpg).

3 Critical Editions, Collations, and other Implementations

In the following section, we present two models for working with textual data in OCHRE: the text corpus model and the critical edition model. In the text corpus model, the researcher creates diplomatic editions of texts in a corpus. Typically, the researcher is establishing a new or updated edition of a text. In the critical edition model, the researcher may be creating text editions, but is also leveraging the graph data model to align various copies of a given text to compare manuscripts and trace transmission variations.

3.1 The Text Corpus Model

First, we illustrate a standard text corpus project and outline some of the most common tasks that take advantage of OCHRE's item-based approach. The *Ras Shamra Tablet Inventory* (RSTI) is an OCHRE research project, based at the Oriental Institute of the University of Chicago. RSTI is directed by the author (Prosser) and Dennis Pardee, professor of Northwest Semitic Philology in the Department of Near Eastern Languages and Civilizations at the University of Chicago. In this project, we organize our research on the culture of Late Bronze Age Ugarit. Ancient Ugarit (modern Ras Shamra) was a city and minor kingdom of the same name in what is now modern-day Syria. The site of Ugarit was occupied almost continuously for nearly six millennia,



Discourse transcription: irrub
Parse: v. G-stem pres. 3rd m. sg.
ir-ru-ub

Figure 4. Dictionary unit for the Akkadian verb *erēbu*, "to enter," with the grammatical form *irrub*, and attested form *ir-ru-ub*.

from the Neolithic Period through to the beginning of the 12^{th} century BCE.[13] Since its discovery in 1929, archaeologists have uncovered thousands of tablets and other inscribed objects. To date, we have assembled a catalog of over five thousand objects in the project.[14] We are in the process of creating digital text editions of the texts recorded on these objects. Also, we have integrated the largest body of digital images of the Ras Shamra tablets.[15] The project serves two purposes: (1) it serves as a central repository for our research, and (2) it provides a mechanism through which we publish our data online. From a practical perspective, in RSTI we make a declarative statement – an interpretive observation – about every grapheme in every text. On this very specific level, we identify the reading of each cuneiform sign, record metadata about the level of preservation, and even write sometimes lengthy prose descriptions to document our epigraphic observations.

The glossary is where we record information about every lexeme attested in our corpus. Every word in a text – a unit of discourse and, conceptually, a node of the graph that represents the text – is associated with a Dictionary unit in a project glossary. A Dictionary unit consists of various grammatical forms, each of which may consist of various attested forms or spellings.

In RSTI, we have two glossaries, one for words in the Akkadian language and one for words in the Ugaritic language, each represented as a hierarchy of lexical entries. The Akkadian texts from Ras Shamra are written in a logosyllabic cuneiform writing system. The Ugaritic texts are, for the most part, written in an alphabetic cuneiform writing system. We transcribe both languages using the Latin script. The writing systems are also OCHRE items, capturing the attributes of different languages and

[13] See Yon et al (1995). Among the many reasons that Ugarit draws the attention of researchers is that the inhabitants of the site produced a fascinating corpus of textual material, from grand mythological tales, to personal letters, down to simple administrative records. See the contributions in Chapter 3 of Watson and Wyatt (1999).

[14] See the project website http://ochre.lib.uchicago.edu/RSTI/teo.html (last accessed September 22, 2020) and blog https://voices.uchicago.edu/rsti/ (last accessed March 17, 2019). For a description of how we have integrated GIS data with the object and text data, see Prosser (2018).

[15] Many of these photos were produced by the epigraphic team of the joint Syro-French Mission de Ras Shamra.

providing a catalog of all valid graphemes and values against which to match and validate textual content. In either system, any given grammatical form of a word may be written in a variety of orthographic forms. Further, there is a high degree of homography in both writing systems. Any given spelling may represent one of many words, or even one of many grammatical forms of those words. The structure of the glossary allows the researcher to disambiguate various word forms.

A word in a text is linked to an attested form in the glossary. Because this attested form is defined as a hierarchical child of a grammatical form, the word in the text inherits the grammatical properties assigned to the grammatical form in the dictionary.[16] In other words, OCHRE's data model represents the attested and grammatical forms as nodes in the graph that captures the relationships among these lexical entries.

Many of these texts are personal correspondence and other types of administrative documents from the royal palace: letters from the king and queen of Ugarit, letters to foreign dignitaries, letters of international intrigue, lists of land distribution to and from named individuals. These texts attest a wide variety of personal names. We are working to identify a prosopography of the persons named in these texts.[17] The graph data model has proven to be a powerful and flexible approach to this research goal. Specifically, the problem is to identify discrete individuals mentioned in texts, what we refer to in OCHRE as *agents*. In the Agent category, we identify an ancient person.[18] Here we can disambiguate various persons who share the same personal name.[19] In the semantics of graph networks, this person is a node in the graph. The many attestations of the person's name in texts are nodes of a discourse unit type. These are linked to the items (nodes) representing the persons being identified as agents. The network extends to include the glossary. Each attestation in a text is linked to a node in the glossary that represents the grammatical form of the name. In the end, we create a network of three nodal hubs: the agent, the textual attestations, and the grammatical entry. This arrangement helps us disambiguate names that are shared by different persons.

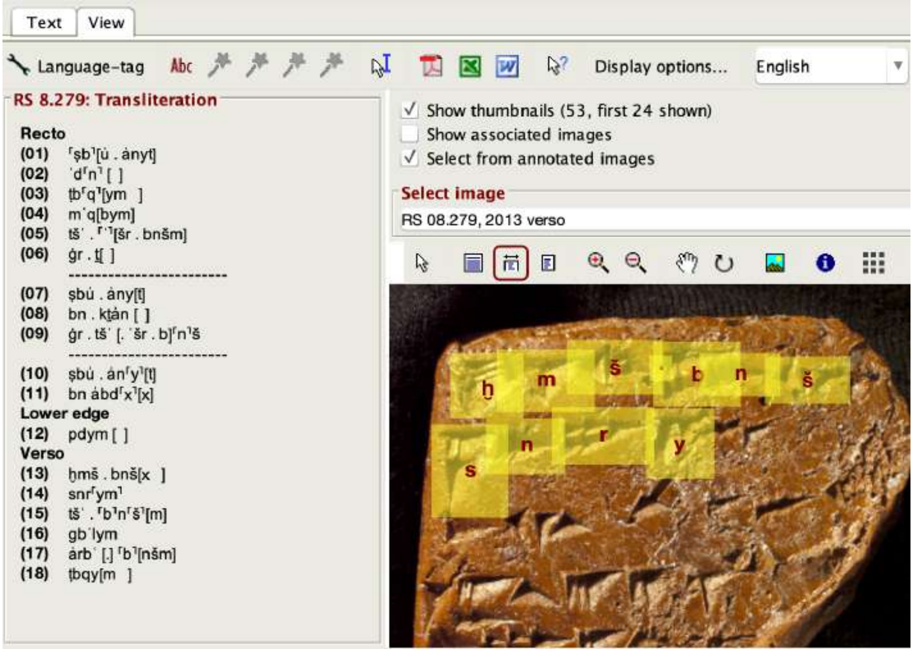
As mentioned above, RSTI includes a large collection of digital images of clay tablets. These images are all accessible through OCHRE, where they are presented alongside views of the text. To further integrate the textual and image data, we use a process called hotspotting to link graphemes to regions of the image where

[16] For a longer discussion of RSTI and its use of OCHRE to perform digital philology, see Prosser (2018), *Digital Philology*.

[17] In this context, we use the term prosopography to refer to the assembled set of familial, occupational, and other information that identifies a person and their relationships to other persons.

[18] See below a discussion of the ontological categories of data in OCHRE.

[19] In our texts, there are many persons identified by a single name only. In this period, there are no surnames, per se. Sometimes persons are listed with a patronymic affiliation, i.e., PN1 son of PN2. However, it is very common for persons to be identified by a single name only.



RS 8.279: Transliteration

Recto
(01) ⸢ṣb⸣[ù . ảnyt]
(02) ʿd⸢n⸣ [ ]
(03) ṯb⸢q⸣[ym]
(04) mʿq[bym]
(05) tšʿ . ⸢ʿ⸣[šr . bnšm]
(06) ġr . ṯ[ ]
----
(07) ṣbú . ảny[t]
(08) bn . kṯán [ ]
(09) ġr . tšʿ [. ʿšr . b]⸢n⸣š
----
(10) ṣbú . ản⸢y⸣[t]
(11) bn ảbd⸢x⸣[x]
Lower edge
(12) pdym [ ]
Verso
(13) ḫmš . bnš[x]
(14) snr⸢ym⸣
(15) tšʿ . ⸢b⸣n⸢š⸣[m]
(16) gbʿlym
(17) ảrbʿ [.] ⸢b⸣[nšm]
(18) ṯbqy[m]

☑ Show thumbnails (53, first 24 shown)
□ Show associated images
☑ Select from annotated images

Select image
RS 08.279, 2013 verso

Figure 5. Image showing hotspot links between transcription and image (a click on ḫ in the image lights up the letter in line 13).

these graphemes are visible. In other words, each image can be marked up with polygons that are linked to epigraphic units in the text. When viewed together in OCHRE, the transcribed text can be synchronized with the hotspot polygons so that a click on a polygon in the image highlights the associated epigraphic unit in the text transcription. This tool has pedagogical utility, but it also useful for clarifying one's interpretation of a damaged sign.

3.2 Text Critical Model

The RSTI text corpus project in OCHRE is representative of numerous other projects that follow this same model: objects, texts, persons, images, analysis, and publication. A different group of text-based projects having different research goals and different source materials falls in a category that we think of as the text critical model. These projects, such as the *Critical Editions for Digital Analysis and Research* project (CEDAR), use the OCHRE platform to perform text criticism and to produce critical editions.[20] From the CEDAR website:

> [t]he goal of the project is to develop, test, and document new methods of digitally representing, displaying, and analyzing manuscripts, textual variants, and diverse editorial readings and translations, enabling views of these data that are not possible using traditional printed editions, with explicit representation of all the intra- and intertextual relationships a scholar may wish to note.[21]

These new methods benefit from reuse and sharing of items made possible by OCHRE's item-based approach, creating somewhat different kinds of graphs as compared to the text corpus projects, but using the same strategies of organizing (hierarchies) and linking of nodes. CEDAR demonstrates that the granular and generic structure of OCHRE suits text critical studies over a wide variety of text corpora. To date, CEDAR includes: (1) the Sumerian editions of the Gilgamesh epic; (2) Hebrew, Greek, Latin, Syriac, and Coptic manuscripts of selected chapters of Genesis, Proverbs, and Daniel; and (3) Shakespeare's Hamlet and Taming of the Shrew. Plans are already underway to expand CEDAR to include additional Biblical books, selections of Sanskrit and Middle Bengali literature, as well as editions of the Egyptian Book of the Dead. Note that text critical projects are an extended use case of the text corpora model, benefiting from all the features described above.

Using OCHRE's item-based approach, the CEDAR project compares textual variants across manuscripts on a letter-by-letter basis. For Biblical scholars interested in text criticism, it is standard to compare a wide variety of manuscripts to investigate the transmission history of a given text. For example, one may wish to compare Medieval Hebrew manuscripts like the Leningrad Codex, Dead Sea Scrolls that attest Biblical passages in Hebrew, Greek manuscripts of the Septuagint, and later Latin and Coptic

[20] The CEDAR project brings together University of Chicago faculty member from various departments: Simeon Chavel (Associate Professor of Hebrew Bible, The Divinity School), Whitney Cox (Associate Professor and Chair, Department of South Asian Languages & Civilizations), Thibaut d'Hubert (Associate Professor, Department of South Asian Languages & Civilizations), Ellen MacKay (Associate Professor, Department of English Language & Literature), David Schloen (Professor of Near Eastern Archaeology, Department of Near Eastern Languages and Civilizations), Jeffrey Stackert (Associate Professor of Hebrew Bible, The Divinity School), and Christopher Woods (John A. Wilson Professor of Sumerian, Department of Near Eastern Languages and Civilizations). The project has benefitted greatly from project personnel: Sarah Yardney (PhD, Divinity), Joseph Cross (PhD Candidate, NELC), Doren Snoek (PhD Candidate, NELC), Andrew Wilent (PhD Candidate, NELC), Ashleigh Cassemere-Stanfield (PhD student, English Language & Literature), Arianna Gass (PhD student, English Language & Literature), Sarah-Gray Lesley (PhD student, English Language & Literature), and Colton Siegmund (PhD student, Near Eastern Languages and Civilizations).

[21] See https://cedar.uchicago.edu/ (Last accessed March 19, 2019).

manuscripts. For these scholars, it is critically important that every letter – and even every accent, vowel mark, or punctuation – is a discrete unit for study.

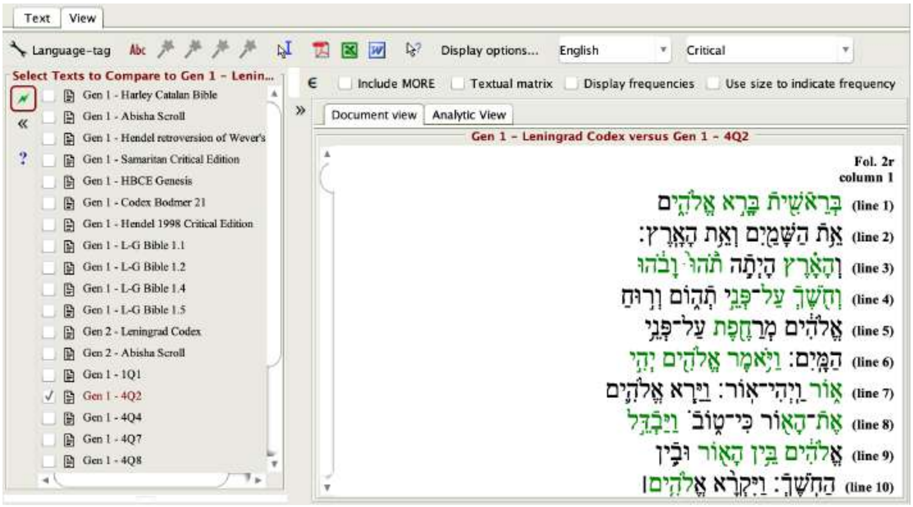
Let's take the Hebrew text of Genesis chapter 1 as an illustrative example. The CEDAR team wishes to compare roughly a dozen Hebrew manuscripts that span about a millennium. Instead of representing each of these Hebrew texts individually, in CEDAR we create a single text that represents all actual content and potential textual variations of the theoretical Hebrew text of Genesis chapter 1. This single text is called a content pool. The content pool is a network of epigraphic units and discourse units, created as a text in exactly the same way as the texts from RSTI described above.

Here we celebrate the power of the item-based data model. Any item in the content pool can be reused, shared among any number of other specific representations of actual texts, what we call local texts. A local text has its own unique epigraphic and discourse hierarchies, but its epigraphic units are borrowed from the content pool. In other words, a user will pick and choose from among the content represented in the content pool, linking in the epigraphic units needed to articulate the structures of the local text. When a given manuscript – i.e., a local text – attests a variant, that variant is added to the content pool and is *aligned with* the non-variant letter by means of a targeted link, thereby extending the pool of available content. To allow for describing variants, adding annotations, or providing extended scholarly commentary, discourse units in local texts are not reused as-is from a content pool, but are auto-aligned with items in the content pool. This arrangement also allows us to compare and align manuscripts across different languages. See more on this below.

From the point of view of the individual nodes, any given epigraphic unit in the graph maintains a list of the necessary edges that identify the texts wherein it appears. For example, the very first letter of the Hebrew text of Genesis chapter 1 is a ב (*bêt*), a preposition meaning "in," as in the phrase, "In the beginning." This epigraphic unit exists in the content pool. The XML document that represents this item lists twelve paths to identify the twelve texts in which it appears. This letter is one node. The twelve texts are each a node. Similarly, each of the over 5,000 epigraphic units in Genesis chapter 1 maintains its own list of the contexts it which it is used, or rather, *reused.*

This systematic reuse of existing textual data whenever possible, eliminates the need for string manipulation when it comes time to compare multiple manuscripts. Texts are automatically aligned by virtue of sharing the exact same database items. Extensive reuse also eliminates the need to proliferate secondary links to make explicit alignments, thereby dramatically reducing redundancy since there is only one copy of the underlying textual content.

As a practical example, for the dozen Hebrew texts of Genesis chapter 1, we do not create a new letter ב (*bêt*) for each phrase, "In the beginning." We reuse the same ב (*bêt*)



Text | View

Language-tag · Display options... · English · Critical

Select Texts to Compare to Gen 1 - Lenin...

- □ Gen 1 - Harley Catalan Bible
- □ Gen 1 - Abisha Scroll
- □ Gen 1 - Hendel retroversion of Wever's
- □ Gen 1 - Samaritan Critical Edition
- □ Gen 1 - HBCE Genesis
- □ Gen 1 - Codex Bodmer 21
- □ Gen 1 - Hendel 1998 Critical Edition
- □ Gen 1 - L-G Bible 1.1
- □ Gen 1 - L-G Bible 1.2
- □ Gen 1 - L-G Bible 1.4
- □ Gen 1 - L-G Bible 1.5
- □ Gen 2 - Leningrad Codex
- □ Gen 2 - Abisha Scroll
- □ Gen 1 - 1Q1
- ✓ Gen 1 - 4Q2
- □ Gen 1 - 4Q4
- □ Gen 1 - 4Q7
- □ Gen 1 - 4Q8

□ Include MORE □ Textual matrix □ Display frequencies □ Use size to indicate frequency

Document view | Analytic View

Gen 1 – Leningrad Codex versus Gen 1 – 4Q2

Fol. 2r
column 1

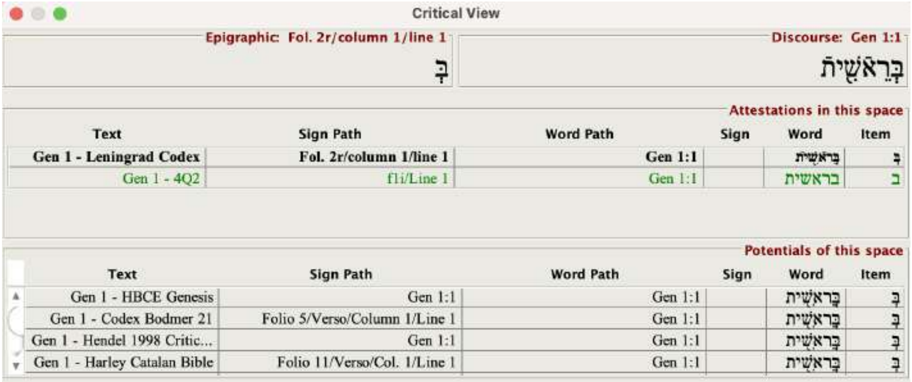
בְּרֵאשִׁ֖ית בָּרָ֣א אֱלֹהִ֑ים (line 1)
אֵ֥ת הַשָּׁמַ֖יִם וְאֵ֥ת הָאָֽרֶץ׃ (line 2)
וְהָאָ֗רֶץ הָיְתָ֥ה תֹ֙הוּ֙ וָבֹ֔הוּ (line 3)
וְחֹ֖שֶׁךְ עַל־פְּנֵ֣י תְה֑וֹם וְר֣וּחַ (line 4)
אֱלֹהִ֔ים מְרַחֶ֖פֶת עַל־פְּנֵ֥י (line 5)
הַמָּֽיִם׃ וַיֹּ֥אמֶר אֱלֹהִ֖ים יְהִ֣י (line 6)
א֑וֹר וַֽיְהִי־אֽוֹר׃ וַיַּ֧רְא אֱלֹהִ֛ים (line 7)
אֶת־הָא֖וֹר כִּי־ט֑וֹב וַיַּבְדֵּ֣ל (line 8)
אֱלֹהִ֔ים בֵּ֥ין הָא֖וֹר וּבֵ֥ין (line 9)
הַחֹֽשֶׁךְ׃ וַיִּקְרָ֨א אֱלֹהִ֤ים (line 10)

Figure 6. Comparing two Hebrew manuscripts of Genesis chapter 1. Green shows agreement, black shows no overlap.

by adding an edge to point to each text where it is attested. Instead of loading multiple texts, then comparing whether twelve different texts all contain the letter ב (*bêt*) in the opening phrase, the OCHRE network generates subsets of the single content pool to represent any specific text. In this way, because all similarities and differences are defined in the underlying data, comparisons do not need to be performed through a secondary process. Any given Hebrew text is simply the selection of nodes that contain the edges that define those nodes as part of the given text. In the opening phrase, "In the beginning," our ב (*bêt*) is the same ב (*bêt*) in every text.

To compare texts across languages, we use cross-cutting links. In a local text, a discourse unit that represents the Hebrew word for "created" is linked to a discourse unit in the Hebrew content pool that represents that same word. This edge creates the discourse relationship between the local text and content pool in Hebrew. The same is true for local texts in other languages, like Greek and Latin. Each discourse unit in those local texts is linked to the appropriate discourse unit in its respective content pool. Creating relationships across languages is done by aligning their content pools. In the Greek content pool, the Greek words for "in the beginning" are aligned with the Hebrew word for "in the beginning" in the Hebrew content pool.[22] Picture, then, the network of words, from a Hebrew local text, through to the Hebrew content pool,

[22] Relationships between discourse units in content pools of different languages can be one-to-one, one-to-many, or many-to-one.



Critical View

Epigraphic: Fol. 2r/column 1/line 1

בְּ

Discourse: Gen 1:1

בְּרֵאשִׁית

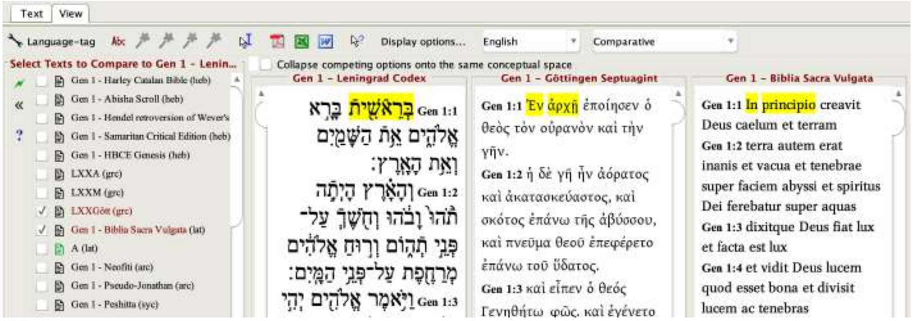
Attestations in this space

Text	Sign Path	Word Path	Sign	Word	Item
Gen 1 - Leningrad Codex	**Fol. 2r/column 1/line 1**	**Gen 1:1**		בְּרֵאשִׁית	בְּ
Gen 1 - 4Q2	f1i/Line 1	Gen 1:1		בראשית	ב

Potentials of this space

Text	Sign Path	Word Path	Sign	Word	Item
Gen 1 - HBCE Genesis	Gen 1:1	Gen 1:1		בְּרֵאשִׁית	בְּ
Gen 1 - Codex Bodmer 21	Folio 5/Verso/Column 1/Line 1	Gen 1:1		בְּרֵאשִׁית	בְּ
Gen 1 - Hendel 1998 Critic...	Gen 1:1	Gen 1:1		בְּרֵאשִׁית	בְּ
Gen 1 - Harley Catalan Bible	Folio 11/Verso/Col. 1/Line 1	Gen 1:1		בְּרֵאשִׁית	בְּ

Figure 7. Details of the first letter of the verse, showing all texts where it is attested.



English

Comparative

Select Texts to Compare to Gen 1 - Lenin...

- □ Gen 1 - Harley Catalan Bible (heb)
- □ Gen 1 - Abisha Scroll (heb)
- □ Gen 1 - Hendel retroversion of Wever's
- □ Gen 1 - Samaritan Critical Edition (heb)
- □ Gen 1 - HBCE Genesis (heb)
- □ LXXA (grc)
- □ LXXM (grc)
- ✓ LXXGött (grc)
- ✓ Gen 1 - Biblia Sacra Vulgata (lat)
- □ A (lat)
- □ Gen 1 - Neofiti (arc)
- □ Gen 1 - Pseudo-Jonathan (arc)
- □ Gen 1 - Peshitta (syc)

□ Collapse competing options onto the same conceptual space

Gen 1 – Leningrad Codex

Gen 1:1 בְּרֵאשִׁית בָּרָא
אֱלֹהִים אֵת הַשָּׁמַיִם
וְאֵת הָאָרֶץ׃
Gen 1:2 וְהָאָרֶץ הָיְתָה
תֹהוּ וָבֹהוּ וְחֹשֶׁךְ עַל־
פְּנֵי תְהוֹם וְרוּחַ אֱלֹהִים
מְרַחֶפֶת עַל־פְּנֵי הַמָּיִם׃
Gen 1:3 וַיֹּאמֶר אֱלֹהִים יְהִי

Gen 1 – Göttingen Septuagint

Gen 1:1 Ἐν ἀρχῇ ἐποίησεν ὁ θεὸς τὸν οὐρανὸν καὶ τὴν γῆν.
Gen 1:2 ἡ δὲ γῆ ἦν ἀόρατος καὶ ἀκατασκεύαστος, καὶ σκότος ἐπάνω τῆς ἀβύσσου, καὶ πνεῦμα θεοῦ ἐπεφέρετο ἐπάνω τοῦ ὕδατος.
Gen 1:3 καὶ εἶπεν ὁ θεός Γενηθήτω φῶς. καὶ ἐγένετο

Gen 1 – Biblia Sacra Vulgata

Gen 1:1 In principio creavit Deus caelum et terram
Gen 1:2 terra autem erat inanis et vacua et tenebrae super faciem abyssi et spiritus Dei ferebatur super aquas
Gen 1:3 dixitque Deus fiat lux et facta est lux
Gen 1:4 et vidit Deus lucem quod esset bona et divisit lucem ac tenebras

Figure 8. Hebrew, Latin, and Greek comparison with a click and automatic highlighting.

then to the Greek content pool, and finally to a Greek local text. This is the network that allows a user to compare local texts across various languages. Note that with a single mapping between the content of one language pool to the content in another language pool, by following the links from the local texts back to the content pool, *any* text in one language can be compared to *any* text in another language using the already-established links between their respective content pools.

4 Implications

The highly granular, item-based, generic model presented by OCHRE for content management will be seen as novel and unfamiliar to many scholars. However, the model is straightforward and easy to understand: items, described by properties, enriched by annotation, organized into hierarchies, and linked to other items. In addition, the undemanding playground of OCHRE's highly generic or *upper* ontology may seem insufficient to the task of competing in the richly-tagged, extensively marked-up world of TEI, or in the universe of the Semantic Web where rival standards – some old, some new – like Dublin Core, FOAF, CIDOC-CRM, FRBR, and so on, vie for adoption, or are constantly being extended. But, allow us to highlight some important implications of OCHRE's approach, not just for textual scholarship, but for scholarship in many academic domains.

First and foremost, the OCHRE approach offers flexibility. The high degree of atomization makes possible the modeling of any type of data. Indulge the following metaphor. If we start with prefabricated walls and floors and roofs, we can build many beautiful and functional structures that conform to the specifications of the original components and the designs of the architects. But if we start with bricks, we can build anything. By atomizing research data to its most minimal meaningful parts, it can be reconstructed in many different ways and for unlimited purposes. The designs are those of the scholar.

The semi-structured, graph data model implemented in OCHRE offers the scholar freedom – freedom not to be locked into a schema designed by someone else or for some other purpose; freedom not to have to decide in advance which schema among many options should be adopted, and then to be trapped in it; freedom not to have to re-tool or transform when standards change or when new best-practices are recommended. OCHRE is, in effect, ontology agnostic. If we need to tag a new concept or feature, we can simply add a new property to the rigorously structured, user-defined taxonomy of our OCHRE project. There is no waiting for an update to an official schema specification to be approved. This is not to say that standards should be ignored. But when using the OCHRE platform, decisions regarding ontological standards are a secondary process, not a primary one. For example, the researcher can map the project-specific taxonomy to the TEI specification and use OCHRE's export or publish function to transform the OCHRE textual items into TEI.[23] Similarly, a different mapping would be used to export OCHRE spatial items to conform to the CIDOC-CRM specification, and so on. A feature to export data as RDF/XML will transform those very same OCHRE items and post them to an RDF triple store. In these transformation processes, OCHRE items, including taxonomic properties, can be mapped onto any number of published ontologies. With tools such as these for exporting and publishing based on semantic mappings, the highly granular OCHRE model gives the user the freedom to play nicely with standards, without first having to store the data in compliance with any single pre-selected ontology.

[23] OCHRE makes it easy to share taxonomies among users or projects, and even to share partial branches of existing taxonomies. In fact, a surprising outcome of the ease of sharing was that, despite OCHRE's extreme flexibility, it had the effect of fostering collaboration rather than encouraging diversity.

The simplicity of the OCHRE data model is matched by the simplicity of a graphical user interface that masks the underlying data format. There are no raw XML files to be edited in oXygen. There is no need to manipulate comma-separated-value (CSV) files. Hundreds, thousands, and even millions of items can be neatly organized into hierarchies within OCHRE's built-in categories of data using a consistent set of tools. The mechanics of the technological tools should not unduly distract the scholar from managing research data.

5 Conclusion

In summary, OCHRE's data model, implemented as a semi-structured graph database, is highly flexible and customizable, yet organized according to a general ontological framework. OCHRE has been used for over a decade now for text corpus projects. Managing approximately 20,000 texts in various languages and writing systems, 200,000 images of texts, and tens of thousands of dictionary entries, the system has proven to be a powerful tool for philological study. As it enters its third year of use for text critical projects, OCHRE continues to evolve. It is the underlying graph data model that makes this recent innovation possible: highly atomized data, organized hierarchically into broad ontological classes, supplemented by cross-cutting linking, all the while supporting reuse wherever possible.

Bibliography

Prosser, Miller, 'Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence', in *CyberResearch on the Ancient Near East and Neighboring Regions: Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*, Digital Biblical Studies (Leiden: Brill, 2018), ii <https://doi.org/10.1163/9789004375086_012>

Prosser, Miller C., 'Digging for Data: A Practical Critique of Digital Archaeology', in "*An Excellent Fortress for His Armies, a Refuge for the People": Egyptological, Archaeological, and Biblical Studies in Honor of James K. Hoffmeier*, ed. by R. E. Averbeck and K. L. Younger (University Park: Penn State University Press, 2020), 309–23

———, 'Digital Philology in the Ras Shamra Tablet Inventory Project: Text Curation through Computational Intelligence', in *CyberResearch on the Ancient Near East and Neighboring Regions (Vol. 2): Evaluating New Tools and Methods for Archaeological Data, Objects, Texts, and Digital Archiving*, ed. by Vanessa Bigot Juloux, Tehri Nurmikko-Fuller, and Sveta Matskevich, Digital Biblical Studies (Leiden: Brill, 2018), 314–335.

Robinson, Iam, Jim Weber, and Emil Eifrem, *Graph Databases: New Opportunities for Connected Data*, 2nd edn (Beijing: O'Reilly, 2015)

Schloen, David, and Sandra Schloen, 'Beyond Gutenberg: Transcending the Document

Paradigm in Digital Humanities', *Digital Humanities Quarterly*, 8.4 (2014) <http://digitalhumanities.org:8081/dhq/vol/8/4/000196/000196.html>

Schloen, J. David, and Sandra Schloen, *OCHRE: An Online Cultural and Historical Research Environment* (Winona Lake, IN: Eisenbrauns, 2012)

TEI Consortium, *P5: Guidelines for Electronic Text Encoding and Interchange*, Version 3.6.0 (TEI Consortium, 2019) <https://tei-c.org/guidelines/>

Watson, Wilfred G. E., and Nicolas Wyatt, *Handbook of Ugaritic Studies*, Handbuch Der Orientalistik, 39 (Leiden: Brill, 1999)

Yon, Marguerite, Maurice Sznycer, and Pierre Bordreuil, eds., *Le Pays d'Ougarit Autour de 1200 Av. J.-C.: Histoire et Archéologie : Actes Du Colloque International, Paris, 28 Juin-1er Juillet 1993*, Ras Shamra-Ougarit, 11 (Paris: Éditions Recherche sur les Civilisations, 1995)

“Standing-off Trees and Graphs”: On the Affordance of Technologies for the Assertive Edition

Georg Vogeler

Abstract

Starting from the observation that the existing models of digital scholarly editions can be expressed in many technologies, this paper goes beyond the simple opposition of ‘XML’ and ‘graph’, It studies the implicit context of the technologies as applied to digital scholarly editions: embedded mark-up in XML/TEI trees, graph representations in RDF, and stand-off annotation as realised in annotation tools widely used for information extraction. It describes the affordances of the encoding methods offered. It takes as a test case the “assertive edition” (Vogeler 2019), in which the text is considered in a double role: as palaeographical and linguistic phenomenon, and as a representation of information. It comes to the conclusion that the affordances of XML help to detect sequential and hierarchical properties of a text, while those of RDF best cover the representation of knowledge as semantic networks of statements. The relationship between them can be expressed by the metaphor of ‘layers’, for which stand-off annotation technologies seem to be best fitted. However, there is no standardised technical formalism to create stand-off annotations beyond graphical tools sharing interface elements. The contribution concludes with the call for the acceptance of the advantages of each technology, and for efforts to be made to discuss the best way to combine these technologies.

1 Introduction

The debate surrounding the best technology stack for digital scholarly edition is ongoing. Recently it has focussed on an opposition between XML and graph technologies. Formally, this opposition does not exist, as you can use XML as a serialisation of graphs (RDF/XML being the best example), and you can express the XML meta-model as a graph (the XML tree is just a rooted and ordered graph). The debate is also a debate between *established* XML users, backed by the wide availability of suitable technologies, and the more recent graph database users[1], experimenting with new solutions, and finding help from a very supportive software company (neo4j.com), for instance. This paper goes beyond the social context and the mathematical models in

[1] See the contribution of Sippl et al. in this volume for an example.

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 73–94.

use. It will explore the implicit contexts of the technologies under debate by examining the metaphors used for the meta-models, the serialisations, and the tools applied to create and process data in scholarly editions. It focusses on one type of scholarly edition (for which I coined the term "assertive edition" (Vogeler 2019), which could also carry labels like *historical edition*, *content-oriented edition*, or *semantic edition*) but will try to transfer observations from this type to other editorial genres.

The assertive edition is a type of scholarly edition that focuses on the content dimension of text in Sahle's text wheel (2013, III, 45–49). The assertive edition tries to represent the information that the authors of the texts want to communicate, or that the readers expect to be communicated, and thus the *real* world described by the text. At the current stage of development, it includes the annotation of terms and named entities, and the addition of descriptive metadata traditionally applied to scholarly editions, while extending editorial practice to add a data layer representing the assertions made by the text. The method has precedents in scholarly editions created by historians. The idea of taking textual documents as information carriers reporting historical facts is, of course, widespread in the context of historical research, but has also been applied to scholarly editions of philosophical texts (Pichler & Zöllner-Weber 2013, Pichler 2020). Indeed, *kleio* databases (Thaller 2003–2009), or relational databases/spreadsheets inserting transcription into column-based data structures, can be considered as early examples of the assertive edition. The development of W3C standards for formal representations of graph-based data models made the method more viable. The assertive editions created at the Zentrum für Informationsmodellierung at Graz University in the context of its humanities research data repository and publication platform *GAMS* (Zentrum für Informationsmodellierung 2014–2020), and those created in the context of *symogih*-infrastructure in Lyon (simogih.org 2012–2020, Beretta 2020) mix representations of the data in RDF with TEI/XML. The question is, is this a good choice? What effect does it have?

The paper discusses the technical solutions to realise assertive editions by the affordances of the technology used. *Affordance* refers to the concepts of James J. Gibson (1977) and Don Norman (1988). Gibson describes the affordance of an object as the possible action of any object, and Norman restricts this to the perceivable actions. I would like to apply it not to a single object, but to a technology, in order to formalise a definition. Thus, the affordances of XML encoding, of RDF formal semantics and triple stores, of Graph databases, of annotation tools etc., are here neither only the theoretical mathematical and computational capabilities of these technologies, nor the human–computer interfaces of the individual implementations. Rather, they refer to the mental models of activities associated with the metaphors dominating the technology in daily scholarly editing practice. The affordance of these technologies can be described as the easily perceivable activities of marking up text, structuring text, connecting entities in a text, expressing knowledge as triples, etc.

Affordance, in this sense, includes prototypical and widely used tools and elements of code, but is not restricted to them.

This follows the pragmatic concept of modelling in the digital humanities forwarded by Arianna Ciula, Øyvind Eide, Cristina Marras and Patrick Sahle (2018, Ciula & Eide 2017). It recognises that epistemological work in the Digital Humanities is often based on external meta-models. This work uses non-computational representations, iteratively translates them into computational implementations, and uses the response to modify the model. This highlights the importance of serialisations and tools for knowledge creation both in and with the model, as they can trigger different metaphors.

Therefore, the affordance of a technology can be considered as the trigger for the selection of a non-technological meta-model. The technical solutions use metaphors like 'hierarchies' (e.g., as 'trees', or 'nested lists'), 'annotations', 'links', 'graphs', 'triples' etc., to describe their meta-model. Which possible use cases are brought to mind by these metaphors? When creating an assertive edition, is it easier to think in terms of hierarchical structure and embedded annotation (XML in an XML editor like Oxygen XML or XMLSpy, and using XPath- and XQuery-based querying), vertex and edge relationships (graph technologies such as RDF in triples stores and labelled property graph databases like neo4j), or separation of base-text and annotation (stand-off annotation tools)?

This approach explicitly mixes data model, serialisations and available tools. The human interaction is with a mixture of the three, and it is human interaction with technology that creates the affordance of the technology.

This paper leaves out several technologies that are theoretically available for digital scholarly editing, but have not gained much acceptance in the scholarly editing community. Relational databases, for instance, are extensively used in software engineering, but are not widely used in the field of scholarly editing, so it makes sense not to discuss affordances of XML-enabled relational SQL database systems, or the implementations of conceptual models using SQL instances. The same is true for scholarly editions based on default content management systems like Drupal or Typo3, which are not used by the wider community of digital scholarly editors. Certainly, it would be worth to study social context of technologies as a reason for their acceptance, but this study focusses on the epistemological implications of the technologies.

The same is true for technologies that were previously used to create scholarly editions, but have fallen out of use — or technologies that were only proposed, but never put into editorial practice. This teaches us again, that social context is one of the things that drive a community to adopt one or the other technology. Sometime, it is hard to distinguish if affordance or social context drives decisions, as affordance, in part, shapes the social context. Manfred Thaller's *kleio* (Thaller 2003–2009), for instance, was a very effective way of representing complex data structures together with the original transcription. However, the programming language used for creating and manipulating these structures did not correspond to the skills taught in computer science introductions, and the software lacked a graphical user interface for a long period. Thus, the emerging digital humanists started their work with tables (and spreadsheet software) and standard relational databases. While SGML, to give another example, had mechanisms to handle overlapping mark-up in concurrent trees, and was therefore well-suited to scholarly editing problems, computer scientists preferred the strict hierarchy of XML. Suggestions for inline mark-up handling overlapping structures like *TexMECS* (Huitfeldt & Sperberg-McQueen 2001), *LMNL* (Tennison & Piez 2002), GODDAG (Sperberg-McQueen & Huitfeld 2004), or *EARMARK* (Peroni 2012–2020; Di Iorio et al. 2009) never offered enough data manipulation possibilities for the technologies to flourish beyond the academic context of proposals and single projects. Excluding these technologies that are not really used in digital scholarly editing reduces the influence of group behaviour in the analysis. It can focus on the perceived affordances of the established technologies, and the metaphorical concepts related to these affordances.

2 XML/TEI

The major technological standard for current editing practice is XML/TEI. For instance, the German research funding scheme DFG this is, for instance, recommended as standard for any scholarly edition (DFG 2015). The digital preservation community accepts XML/TEI as storage format because the definitions of the TEI to encode texts added to the documents gain a semantic explicitness beyond the individual project, and make them fit for digital archiving. Sociologically, the full set of W3C standardised X-technologies offers a well-established technological infrastructure for XML/TEI. The Text Encoding Initiative is probably the largest Digital Humanities semantic data modelling community, and has, since its foundation, focused on creating a terminology as close to the humanities tradition as possible.

Considering the implicit consequences of this technology stack, it is necessary to distinguish between XML and TEI. James Cummings (2018) has argued that the model of the TEI expressed in the TEI Guidelines goes beyond assumptions based on perceived affordance of XML. Indeed, the semantics of the TEI offer so many mechanisms that XML can be considered as merely one possible serialisation of the definitions in the TEI Guidelines. Mathematically, there is no problem in serialising the description of a person in tables instead of in a list of elements nested in the XML/TEI element `<person>`, or to use the TEI parallel segmentation annotation for a critical apparatus to express textual variance as a variant graph as, for instance, is done in CollateX (Dekker & Middell 2011).

The main purpose of the TEI community is to provide interchangeable semantics to the annotation expressed in the tag labels (TEI Consortium 13.2.2020). The affordance in this approach is the transportation of established concepts in the humanities into a computer-processable formal language, the ability to *speak to the computer*. For the assertive edition, this affordance seems to be highly useful, as it offers a substantial range of semantic annotations for text: names of persons (`persName`), places (`placeName`), geographic entities (`geogName`), organisations (`orgName`), physical objects (`objectName`), and structured descriptions for each of them (`person`, `place`, `org`, `object`). Bibliographic items may be identified in the text (`bibl`, `title`) and described with a variety of nested elements (e.g., `author`, `title`, `publisher`, `date`, `textLang`) or in predefined structures (`biblStruct`, `msDesc`). There is mark-up for terminological words (`term`) which can be linked to taxonomies (`taxonomy`). Index terms (`index/term`) can be associated with positions in, or ranges of, text. The editor can reuse the established concepts to identify functions of text.

Thus, the TEI offers the possibility to enrich the text with interpretations of its meaning, using terms close to natural language for the purpose. Still, affordance based on the semantics of natural language can be confusing. In court records, for example, a person could have the role of a witness; however the TEI uses `witness` to encode a textual witness in the context of critical apparatus. For the assertive edition, the main concern regarding the semantics provided by the TEI is whether it fits to the domain of interest in the particular case. The TEI semantics still offer easy-to-grasp solutions for this by providing attributes to define specialisations of existing labels (`@type`) or reference to interpretations (`@ana`).

The affordance of XML is quite different from semantic tagging of entities: XML annotations follow the paradigm of embedded mark-up in a single rooted tree. In the context of scholarly editions, the main consequence of this affordance of XML mark-up is the conceptualisation of text as a sequence of strings, separated into ranges by start and end tags. These text fragments can carry annotations expressed in the labels of the tags. XML implements the basic ideas of the OHCO (Ordered Hierarchy of Content Objects) model of text (Renear et al. 1990), i.e., the metaphor that text can be handled by super-/substructures, and by order assigned to textual fragments, as 'content', As long as you describe texts as a collection of sentences, and each sentence as a collection of words, the OHCO model fits activities necessary for scholarly editing – and, in particular, assertive editions, where sentences and words build basic entities to represent real-world phenomena. Recently, Steven DeRose has summarised the relationship between the OHCO conceptual model of text and XML and concluded "XML is particularly good for documents not because of syntax details, but because its native constructs map readily to document models which have proven useful for serious work with non-ephemeral text documents." (DeRose 2020)

However, the metaphorical potential of the element hierarchy of XML extends the OHCO model. Jennifer Tennison describes the distinction between "containment" as a happenstance relationship between ranges of text and "dominance" as the hierarchical relationship with a meaningful semantic (Tennison 2008). Thus, the nesting of elements can either be just a mereological relationship of containers of several objects, or form a semantic context for the nested elements.

The XML definitions of the TEI semantics make use of the *canonical* order of the XML syntax for the *semantic* order of the textual objects modelled. The ranges defined by the TEI mark-up divide the text into a collection of textual fragments, e.g., sections (`div`), paragraphs (`p`), and referencing strings (`rs`, `name` etc.). These ranges can nest, i.e., one textual fragment can be part of another (containment), and this creates semantic context (dominance). For the assertive edition, semantic context is crucial: a headline to a list, for instance, adds semantics to each entry in the list (Goody 1977, Dolezalova 2009). The XML metaphor fits this need, and, in fact, in XML this context can easily be accessed from each entry in the list by an XPath pointing to the containing list and its heading (`./ancestor::list[1]/head`) or just to the first preceding heading (`./preceding::head[1]`).

Thus, we have two affordances to consider: the affordance of the TEI vocabulary, and the affordance of its serialisation in XML. The main affordance of the X-technologies stack remains the manipulation of OHCO, that is, as nested textual fragments. They can easily be addressed by expressions in XPath, which was designed to navigate the hierarchy and sequence of XML elements. Finally, embedded mark-up creates pointers to text ranges which are described in the annotation, imitating, at least partially, manual annotations in a physical text.

3 Graph Databases

The second major data modelling method in the digital humanities is grounded in graph theory. A graph model of text has been considered by digital scholarly editors (Van Zundert & Andrews 2016; Dekker and Birnbaum 2017). The application of graph models was strongly triggered by its affordances for the representation of variants. A graph model can easily represent paths of alternative readings (Schmidt & Fiormonte 2006; Schmidt and Colomb 2009; Schmidt 2010). However, this use-case does not readily apply to assertive editions.

Andreas Kuczera (2016a, 2016b) has made a case for the graph model from the perspective of an historian. He claims that the graph model makes annotation more flexible. In fact, when all textual entities become identifiable entities, they can recombine in multiple ways: a linguistic fragment can reference single content entities, such as named entities, it can be part of a complex graph, representing assertions, or it can reference the complex graph itself. Having single words as identifiable entities, co-references can be expanded beyond textual sequences ("dieser – jener", "the first, the second,…."): Expressions like "Count Eberhard and his son donate their property in Schmie to the monastery of Maulbronn. The latter promises to add the Wannenwald after the death of his father" (my example), can use expressions of equivalence between the references "the latter" and "his son" (both to be identified as `viaf:80363599`), or "his father" and "Count Eberhard" (both to be identified as `viaf:80337369`). These abstract identifiers can be considered implementations of the conceptual separation between signifier and signified (De Saussure 2013), or between thought, symbol, and referent (Odgen & Richards 1923); or, more simply, support the idea of the assertive edition, that information conveyed by text can be presented separately from the text itself.

The network metaphor lead to very early realisations of assertive editions: digital scholarly editions can enrich named entities in the text by pointing to identifiers from authority files (Poupeau 2006) or in structured data. The *Carl Maria von Weber-Gesamtausgabe* (Allroggen et al. 2011–2019) serves as an example of a 'linked data edition', as it provides a JSON-LD representation of its rich indices on persons, places, letters, works, documents and even commentaries. They reference the idea of the semantic web as a "giant global graph", in which the reader/user of a scholarly edition can drag information from many resources into the edition (Wettlaufer 2018). The assertive edition includes the concept of linked data editions: globally identifiable data points are a good representation of the semantic layer of a text, and the idea of contributing to the general knowledge graph is a valid metaphor for the general purpose of scholarship.

Graph theory implies an affordance which has made it highly attractive for social analysis: network structures can be visualised as graphs. Edges as links between ideas were connected to associative thinking in Deleuze and Guattari's "rhizome" image of thought (1980). One main attraction of graph technologies comes from visualisations that allow jumping from one node to the next. Force-directed graph drawing methods (Eades 1984; Fruchterman & Reingold 1991) create visual impressions of groups with high interconnection, and give an easy overview of the organisation of a graph. This kind of affordance is used when conceptualising textual relationships as nets (Andrews & Macé 2013; Andrews & Van Zundert 2012). As every social network can be expressed as a graph, the meta-model supports the historian's interest in people connected with one another. This affordance, at the very least, makes graph technologies attractive for the implementation of assertive editions.

Graph-based technologies are linked also to knowledge representation in semantic networks (Quillian 1967). James Hendler demonstrated that a substantial part of description logic can be expressed with an RDF-based vocabulary (Hendler et al. 2005; Bechhofer et al. 2000). Based on this, the Web Ontology Language (OWL) (Lacy 2005, 134; W3C OWL Working Group 2012) was created. It is a vocabulary that realises graphs that express the logic of OWL-DL. In encoding practice, the class hierarchies, which form the entry screen to the main OWL editor Protégé, are often taken as the main affordance of OWL. In fact, the formal affordance of OWL is not really exploited in Digital Scholarly Editions, or even in Digital Humanities at all. A practical effort to define digital editing-related concepts as a formal ontology is made by the Swiss *NIE-INE* project (*NIE-INE* 2020)[2].

In the context of the assertive edition, a special type of graph comes into play: the sentence. The W3C Resource Description Framework (RDF) adds this metaphor to a basic graph model: 'subject predicate object' structures translate easily into directed graphs (W3C 2014). In promoting the semantic web, this metaphor helps one to talk about the data formalised in RDF. The metaphor even leads to suggestions like that by Roland Kamzelak (2016) of taking an approach to creating RDF triples that is based more on natural language. This suggestion is close to the affordance of controlled natural languages like Attempto Controlled English (ACE) (Fuchs et al. 2006; Fuchs 2018), which are used for knowledge representation. Alexandr Ivanovs and Alexey Varfolomeyev (2014), for instance, have used ACE experimentally in a scholarly edition of charters.

The W3C has realised that this metahpor can be misleading: the label "semantic web" contributed to an unnecessary combination of the concepts of "linked data" with "knowledge representation". W3C has changed the label for its activities in the field since 2013 to "Web of Data". In fact, RDF can be used to model hierarchical knowledge systems as well: The Simple Knowledge Organisation System (SKOS) recommended by the W3C (W3C 2009) proposes a set of broader and narrower relationships which are the major type of relationship in most use cases of SKOS.

The triple-based sentence metaphor of RDF has drawbacks. Firstly, simple 'subject predicate object' sentences do not suffice to express the context of propositions such as *The people of London swear allegiance to King George I. – in 1723* (Vallance 2013) and *The city of Basel received 31 and a half pounds on wine tax – in Füllinsdorf, Lupsingen and Zieten* (Burghartz et al. 2015), which add temporal or geographical constraints to the proposition. The standard RDF solutions for these extended sentences are blank nodes (W3C Working Group 2014, #section-blank-node) and property singletons (Nguyen et al. 2014; Nguyen et al. 2015). Both methods are much less intuitive than labelled property graph technologies, e.g., Apache TinkerPop (Apache Software Foundation 2015–2019) or Neo4j (Neo4j 2010–2020), which allow the description of edges with properties.

[2] See also the contribution by Cools & Padlina in this volume.

Therefore, the major affordance of graph databases in an assertive edition is the use of a semantic network to connect pieces of information. The specifications of RDF add the metaphor of a 'statement' to express data as simple propositions. The W3C web-of-data stack adds to this the creation of abstract identifiers, helping to separate text and data, and to prepare the data for integration into a global network of information.

4 Stand-Off-Annotation

4.1 Layers

The assertive edition has a multi-layered approach to text. From the theoretical point of view, this need is well explained by Börje Langefors' "Infological equation" (1966): information is a function of time and data. The content of the edited text is information extracted by the editor from the base text under specific conditions. Editorial annotation and formal representation of scholarly readings of the text should be separated from the text, as Manfred Thaller has pointed out (Thaller 2012). In fact, scholarly discourse on the content of the text of an edition is very often triggered by the attempt to avoid combining scholarly interpretations with documented physical and graphical observations on the manuscripts (Zeller 1971).

There are a number of technologies dedicated to *stand-off annotation*. The separation of primary data and mark-up by semantic links dates back into the 1990s (Thompson & McKelvie 1997) and is the de facto standard in the annotation of image, audio and video material, mainly because the encoding of the annotated data is significantly different from that of the annotation (literals). Stand-off annotation has been applied to linguistic annotation of texts for a long time. Several approaches to encoding this annotation are in use: tabular lists of tokens (e.g., *TCF* Weblicht 2015), pointers to offsets in the text stream (e.g., *PAULA*, Zeldes et al. 2013), or pointers to other anchors in text (ISO-LAF 2012; Ide & Sudermann 2014) such as, for instance, embedded mark-up. The general affordance of stand-off annotation methods for assertive editions lies, on the one hand, in the effective processing of the annotations themselves, as they become separate objects with own properties, and, on the other, in the layered semantic conceptualisation of the text. Here, I want to argue that the combined affordances of XML and RDF are not sufficient for this layered approach.

Stand-off annotation is very often introduced as a solution to the problem of overlapping mark-up in XML. This opposition is informative for the perception of affordances of the technologies. Many stand-off annotation formats use XML for serialisation, and the Guidelines of the TEI have a full section describing how to create stand-off mark-up with the TEI (TEI Consortium 2020b; Cummings 2018). Still, XML is perceived as a strict hierarchy.

As explained above, the implicit semantics of XML attributes support a layered approach. Syntactically, attributes are only one type of nested node. The standard semantics assigned to this type of substructure supports this multi-layered approach: The distinction between the three types of nodes (attributes, sub-elements, and content) suggests that attributes are a separate information layer to the annotated text. In fact, attributes in the TEI can create alternative representations (e.g., `date@when`, `num/@value`, `measure/@unit|@commodity|@quantity`), and this fits in very well with the concept of the assertive edition. The generic `@ana` attribute allows these to be extended to include a layer that is completely defined by the editor.

However, the TEI does not insist on the layer metaphor for attributes: on the one hand, nesting of elements can also create layers of text, e.g., the `choice` or the `app`-element expressing alternative representations of one text. On the other hand, attributes suggest an *isomorphism* metaphor, when they encode specialisations (`@type`), or a network metaphor when they encode references (`@ana`, `@ref`, `@target`, `@spanTo`, `@facs`). This last affordance, i.e., the reference from names or reference strings to formal descriptions of the entities is, as explained above, very close to the needs of the assertive edition, but would be used rather in the context of the layer metaphor than as a network.

This is not a critique of the XML implementation of the TEI, but demonstrates that the metaphors applicable to XML's syntax might be stronger than the semantics proposed by the TEI. The expectation that nesting should be more than just an issue of serialisation, that it should have a specific meaning, drives well-known discussions of the type "why can't element X contain element Y?" In fact, the TEI goes beyond the primary affordance of XML, when it breaks the logic of embedded mark-up: semantic annotation in the TEI, for instance, uses pointers from a linguistic fragment to form a list-like data structure: `persName` points via `@ref` to `person`, for example, and `listPerson/person` describes the person as a list of properties. TEI has introduced other constructs to handle typical drawbacks of XML, like the `@part` attribute for overlapping mark-up, and a more generic method to create a sequence of XML elements beyond the sequence in the document (`@prev`, `@next`).

The distance between the technologies used to process XML and the formal openness of the TEI can be demonstrated in the handling of overlapping mark-up. TEI proposes `@part=I|M|F` or `@prev` and `@next`. Both need complex XSLT or XQuery expressions to create the merged node (listing 1a and 1b). By contrast, stand-off technologies could just use a single `name[offset-start, offset-end]` expression, if the stand-off annotation were expressed by offsets in the range of the basic text.

The same is true for stand-off annotations expressed in XML/TEI, which need to resolve the ID-references provided in `@ref` or `@spanFrom/@spanTo`.

The same observation can be made when expressing layered information with graph technologies: RDF offers the reification vocabulary to express that statements are made by a person about a subject. Reified statements can express a single interpretation of a given text, and also the fact that this interpretation might be different from interpretations by other scholars. Reification proposes the creation of a statement graph (`rdf:type rdf:Statement`) that is composed of triples describing the role of entities and literals in a statement (`rdf:subject`, `rdf:predicate`, `rdf:object`). The mathematical affordance of RDF is sufficient to model the problem, but the solution does not meet the intuitive needs of scholarly editors. W3C has introduced named graphs as an alternative solution. Named graphs use an IRI to identify a full graph consisting of many triples (Carroll et al. 2005; Bizer & Cyganiak 2014). The method fits much better with the everyday experience of receiving RDF as document on the web via a specific URL.

```
<xsl:template match="*[@next]">
  <xsl:element name="name()">
    <xsl:apply-templates/>
    <xsl:apply-templates select="following::*[@xml:id=current()/@next"/>
  </xsl:element>
</xsl:template>

<xsl:template match="*[@part='I']">
  <xsl:element name="name()">
    <xsl:apply-templates/>
    <xsl:apply-templates select="intersect(current()/following::*[name()=current
         ()/name()][@part='F'][1]/preceding::*[name(), current()/name()][@part='M
         '])"/>
    <xsl:apply-templates select="current()/following::*[name()=current()/name()][
         @part='F'][1]"/>
  </xsl:element>
</xsl:template>
```

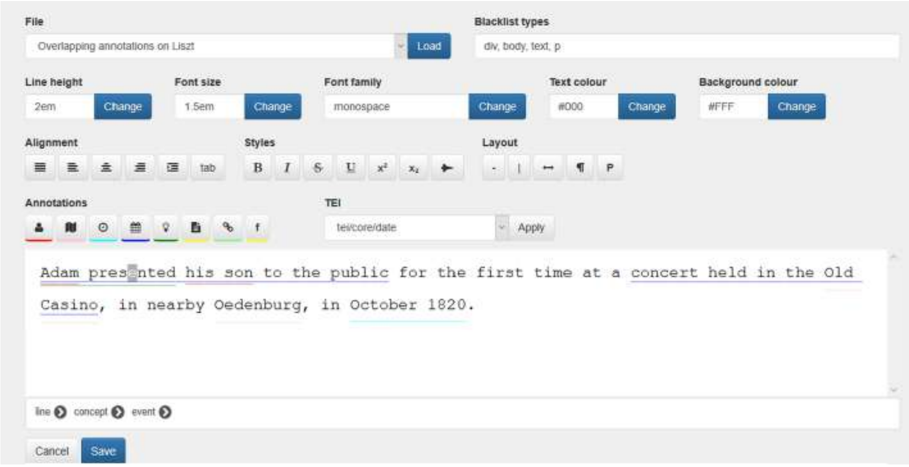
Listing 1. sample XSLT code merging TEI encoding for overlapping mark-up: a) expressed via `@next` pointer, b) expressed via `@part`

An alternative approach is to define a vocabulary with semantics dedicated to the layer metaphor: an XML element `<layer>` can create the necessary affordance, and the recently introduced `<standOff>` element in the TEI might serve this purpose. In RDF, the W3C web annotation data model (Sanderson et al. 2017) and ISO 24612:2012 (ISO-LAF 2012) describe stand-off annotations in a more generic way in RDF.

4.2 Tools

There are a range of tools supporting the layered approach. The definition of the web annotation model by the W3C is rich, but implementations in generic web annotation tools like hypothes.is (hypothes.is [2011–2020]) create only plain text annotation.

Other stand-off annotation tools focus on overlapping mark-up. Catma (Petris et al. 2008–2020), for instance, is extensively used in digital philology (Petris et al. 2008–2020, /publications). It foregrounds the basic idea of annotation by using coloured underlines linked to tag sets. The TEI export of this annotation (Petris


File: Overlapping annotations on Liszt

Blacklist types: div, body, text, p

Line height: 2em

Font size: 1.5em

Font family: monospace

Text colour: #000

Background colour: #FFF

TEI: tei/core/date

Adam pres nted his son to the public for the first time at a concert held in the Old Casino, in nearby Oedenburg, in October 1820.

Figure 1. Screenshot from the SPEEDy annotation editor.

et al. 2008–2020, /documentation/tei-export-format/) creates a TEI structure with pointers between text, empty `<seg>` elements, and feature structure declarations (`<fs>`) referenced by the `@ana` attribute.

A similar approach is taken by the annotation tool “SPEEDy” (Figure 1) developed by Iian Neill in the context of *Codex-net* (Neill 2013–2019).[3] It creates rich stand-off annotations, which are exportable in JSON (argimenes 2018–2020). The editor is still close to the surface of the text, and it focuses on allowing overlapping mark-up (visualised as coloured underlines). However, the abstract concept of this tool is different. It considers annotations as *claims* about text (or other entities). Thus, there is the possibility of creating more complex data structures.

Still, in the context of the assertive edition, the current affordance of the ‘overlapping-mark-up’ approach of Catma and SPEEDy does not extend much beyond embedded mark-up: the annotation is defined by generic tag sets, not by references to individuals.

SPEEDy’s ‘claim’ approach is similar to the factoid model proposed by John Bradley and Harold Short for prosopographical databases (2005; Pasin & Bradley 2015). It models a tripartite data structure, in which the text is only a source of information on a person. The factoid model has been applied to several applications, though typically hidden in the database model, and usually away from the user, behind graphical user interfaces, for instance in the *Personendaten-Repositorium* of the Berlin-Brandenburg

[3] See also the contribution by Neill & Schmidt in this volume.

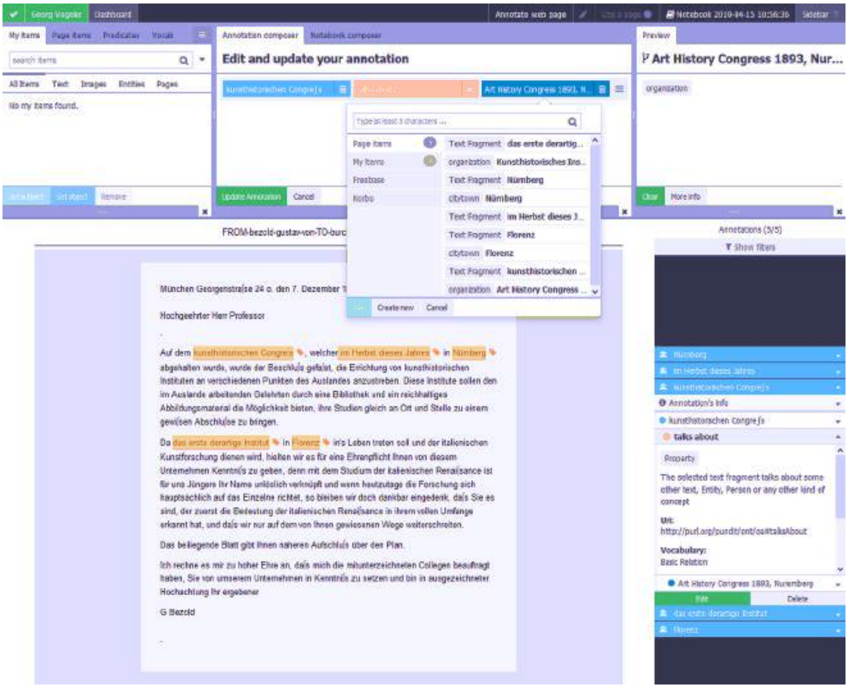
Academy of Sciences (BBAW 2009; Neumann et al. 2011) and its *Archiveditor* (BBAW 2011).

The Archiveditor creates the semantic annotations needed in an assertive edition. Oxygen plugins such as ediarum (BBAW [2014–2019]; Dumont & Fechner 2014) offer similar affordance in the XML/TEI technology stack: users utilise graphical interface elements to link text ranges in TEI encoding to controlled vocabularies or to lists of entities. Users can switch between the XML representation of the text and separate XML documents for the annotations.

The scholarly edition of the letters of Jakob Burkhardt (Ghelardi & European Research Advanced Grant Project EUROCORR 2019) demonstrates semantic annotation using the RDF stack, which meets the needs of assertive editions. For semantic annotation, it used thepund.it, by Net7 (Net7 [2015–2020]), which creates annotations in the form of RDF-based triples (Morbidoni & Piccioli 2015). It allowed for the insertion of text fragments into triples, linking them to other text fragments or concepts defined by the user (Figure 2).

Using stand-off annotation tools to represent relational semantics is a good fit for the use case of the assertive edition: when applying automatic information extraction methods, the identification of named entities, and their position in syntactic structure, creates a different layer of information from the information usually used in databases. Only the extraction of "Who did What to Whom, and When, Where, and How?" converts the text into propositions in a formal structure, and this does not have to be close to the original text. This conversion includes phenomena like co-reference resolution and entity identification, where the linguistic surface cannot be used as identifiers in a semantic representation.

There are several tools for the annotation of semantic relations, and their affordance leads to a common meta-model beyond the representation of overlapping markup. The *BRAT* annotation tool (brat contributors 2010–2018), for instance, has an easy-to-use graphical user interface (see Figure 3), in which the user can identify entities of interest, and link them together. While text fragments represent the entities, label arrows indicate the relationships between the entities. The very similar design of the annotation tool in *Recogito* (Pelagios Network 2014–2020) (see Figure 4) demonstrates a recognition of the basic affordance of stand-off mark-up to be able to create links between references to named entities in a text. They both use flat text as a reference point. *Recogito* exports it to a combination of file-types: RDF files, using web annotation vocabulary; TEI files, with mark-up for the main entity types; and, as CSV lists of nodes and edges, prepared for *Gephi* to represent the relations between the entities. Brat exports as a list of numbered tokens and plain text. Even the far more feature-rich linguistic annotation platform *inception* (de Castilho et al. 2018–2020, https://inception-project.github.io/) makes use of the *BRAT* tool for its annotation.


Georg Vogeler
Dashboard
Annotate web page
Notebook 2018-04-15 10:56:36
Sidebar
My items
Page items
Predicates
Vocab
Annotation composer
Notebook composer
Preview
Edit and update your annotation
Art History Congress 1893, Nur...
organization
kunsthistorischen Congreſs
Art History Congress 1893, N...
Type at least 3 characters ...
Page items
My items
Freebase
Korbo
Text Fragment das erste derartig...
organization Kunsthistorisches Ins...
Text Fragment Nürnberg
city/town Nürnberg
Text Fragment im Herbst dieses J...
Text Fragment Florenz
city/town Florenz
Text Fragment kunsthistorischen ...
organization Art History Congress ...
Create new
Cancel
All items
Text
Images
Entities
Pages
No my items found.
Update Annotation
Clear
More info
FROM-bezold-gustav-von-TO-burc
Annotations (5/5)
Show filters
Nürnberg
im Herbst dieses Jahres
kunsthistorischen Congreſs
Annotation's info
kunsthistorischen Congreſs
talks about
Property
Url:
Vocabulary:
Basic Relation
Art History Congress 1893, Nuremberg
Edit
Delete
das erste derartige Institut
Florenz

The selected text fragment talks about some other text, Entity, Person or any other kind of concept

Url: http://purl.org/pundit/ont/oa#talksAbout

München Georgenstraſse 24 o. den 7. Dezember 1

Hochgeehrter Herr Professor

Auf dem kunsthistorischen Congreſs, welcher im Herbst dieses Jahres in Nürnberg abgehalten wurde, wurde der Beschluſs gefaſst, die Errichtung von kunsthistorischen Instituten an verschiedenen Punkten des Auslandes anzustreben. Diese Institute sollen den im Auslande arbeitenden Gelehrten durch eine Bibliothek und ein reichhaltiges Abbildungsmaterial die Möglichkeit bieten, ihre Studien gleich an Ort und Stelle zu einem gewiſsen Abschluſse zu bringen.

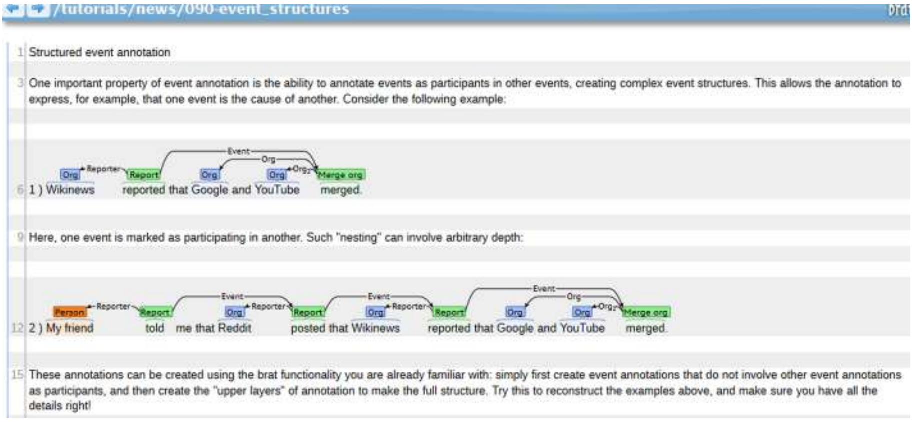
Da das erste derartige Institut in Florenz in's Leben treten soll und der italienischen Kunstforschung dienen wird, halten wir es für eine Ehrenpflicht Ihnen von diesem Unternehmen Kenntniſs zu geben, denn mit dem Studium der italienischen Renaiſsance ist für uns Jüngere Ihr Name unlöslich verknüpft und wenn heutzutage die Forschung sich hauptsächlich auf das Einzelne richtet, so bleiben wir doch dankbar eingedenk, daſs Sie es sind, der zuerst die Bedeutung der italienischen Renaiſsance in ihrem vollen Umfange erkannt hat, und daſs wir nur auf dem von Ihnen gewiesenen Wege weiterschreiten.

Das beiliegende Blatt gibt Ihnen näheren Aufschluſs über den Plan.

Ich rechne es mir zu hoher Ehre an, daſs mich die mitunterzeichneten Collegen beauftragt haben, Sie von unserem Unternehmen in Kenntniſs zu setzen und bin in ausgezeichneter Hochachtung Ihr ergebener

G Bezold

Figure 2. Triples for https://burckhardtsource.org/letter/273 as annotations with thepund.it in burckhardtsource.org/. Screenshot from 2018, in August 2021 the functionality was disabled.



Structured event annotation

One important property of event annotation is the ability to annotate events as participants in other events, creating complex event structures. This allows the annotation to express, for example, that one event is the cause of another. Consider the following example:

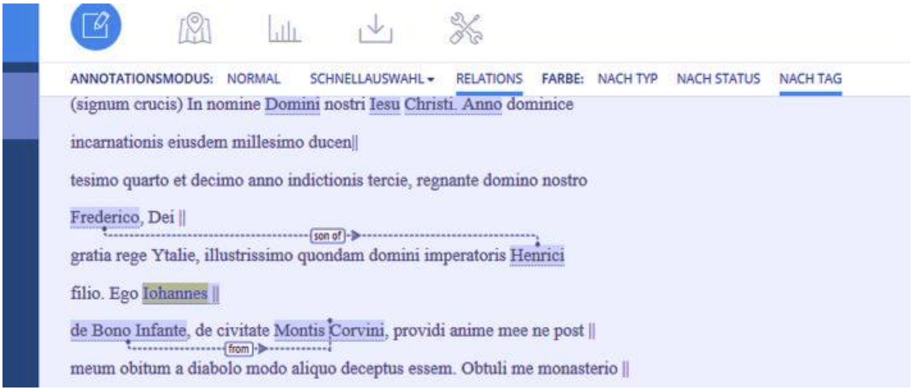
1) Wikinews reported that Google and YouTube merged.

Here, one event is marked as participating in another. Such "nesting" can involve arbitrary depth:

2) My friend told me that Reddit posted that Wikinews reported that Google and YouTube merged.

These annotations can be created using the brat functionality you are already familiar with: simply first create event annotations that do not involve other event annotations as participants, and then create the "upper layers" of annotation to make the full structure. Try this to reconstruct the examples above, and make sure you have all the details right!

Figure 3. Stand-off annotation with *BRAT*. While *BRAT* itself is out of development, many features have been integrated into *inception*.



(signum crucis) In nomine Domini nostri Iesu Christi. Anno dominice
incarnationis eiusdem millesimo ducen||
tesimo quarto et decimo anno indictionis tercie, regnante domino nostro
Frederico, Dei ||
gratia rege Ytalie, illustrissimo quondam domini imperatoris Henrici
filio. Ego Iohannes ||
de Bono Infante, de civitate Montis Corvini, providi anime mee ne post ||
meum obitum a diabolo modo aliquo deceptus essem. Obtuli me monasterio ||

Figure 4. Stand-off annotation with *Recogito.*

To summarise the current state of research, the only technologies that provide explicit affordance for stand-off annotation are graphical tools. For the end user this is no problem. For technically informed users it is less comfortable: even if the design language for *BRAT* and *Recogito* is highly similar, the semantics of the stored data structures can vary significantly, as the brief overview of formats to store stand-off annotations above has shown. There is no encoding standard for stand-off mark-up that is as easy to grasp and to manipulate as pointy brackets in XML, mark-up languages that allowing overlapping mark-up like *TAGML* (Dekker et al. 2018), or RDF triples expressed in Turtle.

5 Conclusion and Future Work

In the considerations above, I have identified the *sequence*, the *hierarchy*, the *statement*, the *network*, and the *layer* as metaphors that describe the interactions of the editor with the data in the creation of an assertive edition. Of course, other edition types might prefer different metaphors. The tree is the basic metaphor of classical stemmatology. Recently, the graph metaphor has been particularly successful in the analysis of textual variants, as it can describe the complexity of diverging text sequences that do not produce a hierarchy. Genetic editions build on generative metaphors like 'parents', 'derived from', or on temporal sequences ('protograph', 'apograph'). Documentary editions have a slight tendency towards a *topological* metaphor, positioning texts in a two-dimensional space, and often realised in stand-off annotations as the standard serialisations of the visual representation (as a matrix cannot easily be inserted into hierarchical models). When Scholarly editing is taken as a basis for linguistic or philological analysis, it shares the tendency towards stand-off solutions, as this analysis is often conceptualised as a multi-layer annotation.

The affordances of the existing technologies as serialisations, as tools, as well as the conceptual meta-models do not support all of these metaphors in the same way. The editor will therefore select technologies better adapted to single tasks, and combine as many of the technologies as possible. The debate about the best technology stack should therefore move towards a debate on the best method for a given combination. To facilitate the interchange of data we should take care to avoid implicit semantics, for instance, by making relationships in TEI that result from the XML meta-model explicit. Effort should be put into the development of formal procedures for converting one serialisation into another, or into making as many data formats as possible available to a single tool, designed for a specific task, without losing the expressiveness of the original data. The DH community should indeed consider standing-off the idea that trees and graphs are fundamentally in opposition to one another. It should consider them rather as metaphors more helpful for one scholarly editing task than for others.

Bibliography

Allroggen, Gerhard, Markus Bandur, Frank Ziegler, Joachim Veit, Peter Stadler, Eveline Bartlitz, Dagmar Beck, and Solveig Schreiter (eds.), *Carl Maria von Weber-Gesamtausgabe*, 2011–2019 <https://weber-gesamtausgabe.de/>

Andrews, Tara L, and C. Macé, 'Beyond the tree of texts : Building an empirical model of scribal variation through graph analysis of texts and stemmata', *Literary and Linguistic Computing* 28,4 (2013), 504–521

Andrews, Tara L, and Joris van Zundert, *Stemmaweb: Interoperable tools for the investigation of medieval text stemmatology,* 2012 <https://stemmaweb.net/>

Apache Software Foundation, 'Apache TinkerPop', 2015–2019 <http://tinkerpop.apache.org/>

argimenes. A Standoff Properties Editor in JavaScript, 2018–2020 <https://github.com/argimenes/standoff-properties-editor>

BBAW, 'Archiv-Editor', Personendaten-Repositorium. DFG-Projekt, 2011 <http://pdr.bbaw.de/software/ae/>

———, 'ediarum - Digitale Arbeits- und Publikationsumgebung für Editionsvorhaben', 2014–2019 <http://www.bbaw.de/telota/software/ediarum>

———, 'Personendaten-Repositorium', Personendaten-Repositorium. DFG-Projekt, 2009 <https://doi.org/pdr.bbaw.de/>

Bechhofer, Sean, Jeen Broekstra, Stefan Decker, Michael Erdmann, Dieter Fensel, Carole Goble, Frank van Harmelen, et al., 'An informal description of OIL-Core and Standard OIL: a layered proposal for DAML-O', 2000 <http://web.archive.org/web/20090320081902/http://www.ontoknowledge.org:80/oil/downl/dialects.pdf>

Beretta, Francesco, 'A Challenge for Historical Research: Making Data FAIR Using a Collaborative Ontology Management Environment (OntoME)', *Semantic Web Journal*, 2020

<http://semantic-web-journal.net/content/challenge-historical-research-making-data-fair-using-collaborative-ontology-management-0>

Bizer, Chris, and Richard Cyganiak, 'RDF 1.1 TriG', W3C Recommendation, 25. Feb. 2014 <https://www.w3.org/TR/2014/REC-trig-20140225/>

Bradley, John, and Harold Short, 'Texts into Databases: The Evolving Field of New-Style Prosopography', *Literary and Linguistic Computing* 20, Suppl. (2005), 3–24 <https://doi.org/10.1093/llc/fqi022>

brat contributors, 'Brat Standoff Format', brat, 2010–2018 <http://brat.nlplab.org/standoff.html>

Burghartz, Susanna, Sonia Calvi, and Georg Vogeler, 'Urfehdebücher der Stadt Basel – digitale Edition', Digital edition. Urfehdebücher der Stadt Basel – digitale Edition (Graz: Zentrum für Informationsmodellierung 2017) <http://gams.uni-graz.at/context:ufbas>

Carroll, Jeremy J, Christian Bizer, Pat Hayes, and Patrick Stickler, 'Named Graphs, Provenance and Trust', in *Proceedings of the 14th International Conference on World Wide Web - WWW '05*, 613 (Chiba, Japan: ACM Press, 2005) <https://doi.org/10.1145/1060745.1060835>

Castilho, Richard Eckart de, Iryna Gurevych, Anna-Felicitas Hausmann, Jan-Christoph Klie, and Ute Winchenbach, 'INCEpTION', Annotation platform. INCEpTION, 2018–2020 <https://inception-project.github.io/>

Ciula, Arianna, and Øyvind Eide, 'Modelling in Digital Humanities: Signs in Context', *Digital Scholarship in the Humanities* 32, Suppl._1 (2017), i33–46 <https://doi.org/10.1093/llc/fqw045>

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle, 'Modelling: Thinking in Practice. An Introduction', *Historical Social Research, Supplement* 31 (2018), 7–29 <https://doi.org/10.12759/hsr.suppl.31.2018.7–29>

Cummings, James, 'A World of Difference: Myths and Misconceptions about the TEI', *Digital Scholarship in the Humanities*, 24, suppl_1 (2019), i58–i79 <https://doi.org/10.1093/llc/fqy071>

De Saussure, Ferdinand, *Cours de linguistique générale: zweisprachige Ausgabe französisch-deutsch mit Einleitung, Anmerkungen und Kommentar von Peter Wunderli* (Tübingen: Narr, 2013)

Dekker, Ronald, and David J. Birnbaum, 'It's more than just overlap: Text As Graph', Washington, DC, 2017 <https://doi.org/10.4242/BalisageVol19.Dekker01>

Dekker, Ronald, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom, and David J. Birnbaum, 'TAGML: A Markup Language of Many Dimensions', in *Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies*, vol. 21 (2018) <https://doi.org/10.4242/BalisageVol21.HaentjensDekker01>

Dekker, Ronald, and Gregor Middell, 'Computer-Supported Collation with CollateX', in *Supporting Digital Humanities 2011: Answering the unaskable, Kopenhagen 17–18 November 2011, proceedings ed. by B. Maegaard* (2011)

Deleuze, Gilles, and Félix Guattari, *A Thousand Plateaus. Trans. Brian Massumi. London and New York: Continuum, 2004. Vol. 2 of Capitalism and Schizophrenia. 2 vols. 1972–1980. Trans. of Mille Plateaux* (Paris: Les Editions de Minuit, 1980)

DeRose, Steven, ‘What is a diagram, really?’, in *Balisage: The Markup Conference 2020, Washington, DC*, 2020 <https://doi.org/10.4242/BalisageVol25.DeRose01>

DFG, ‘Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft’, November 2015 <https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literatur wissenschaft.pdf>

Di Iorio, Angelo, Silvio Peroni, and Fabio Vitali, ‘Towards Markup Support for Full GODDAGs and beyond: The EARMARK Approach’, in *Balisage: The Markup Conference 2009, Montréal, Canada*, 2009 <https://doi.org/10.4242/BalisageVol3.Peroni01>

Digitale Akademie der Mainzer Akademie der Wissenschaften und der Literatur, ‘eXGraphs’, LOD.Academy, 2019 <https://lod.academy/site/tools/digicademy/exgraphs>

———, ‘XTriples’, Webservice. XTriples, 2012–2019 <https://xtriples.lod.academy/index.html>

Dolezalova, Lucie, ed., *The Charm of a List: From the Sumerians to Computerised Data Processing* (Newcastle upon Tyne: Cambridge Scholars Publishing, 2009)

Dumont, Stefan, and Martin Fechner, ‘Bridging the Gap: Greater Usability for TEI Encoding’, *Journal of the Text Encoding Initiative*, Issue 8 (2014) <https://doi.org/10.4000/jtei.1242>

Eades, Peter, ‘A Heuristic for Graph Drawing’, *Congressus Numerantium* 42,11 (1984), 149–160

Fruchterman, Thomas M.J, and Edward M. Reingold, ‘Graph drawing by force-directed placement’, *Software: Practice and Experience* 21,11 (1991), 129–1164 <https://doi.org/10.1002/spe.4380211102>

Fuchs, Norbert E, ‘Understanding Texts in Attempto Controlled English’, in *Proceedings of the 6th International Workshop on Controlled Natural Language (CNL 2018), Maynooth, Ireland*, 2018 <http://attempto.ifi.uzh.ch/site/pubs/papers/cnl2018main_fuchs.pdf>

Fuchs, Norbert E, Kaarel Kaljurand, and Gerold Schneider, ‘Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces’, in *Proceedings of 19th International Florida Artificial Intelligence Research Society Conference*, 2006 <https://doi.org/10.1.1.541.9039>

Ghelardi, Maurizio, and European Research Advanced Grant Project EUROCORR, ‘J.Burckhardt’, Digital scholarly edition. J.Burckhardt, 2019 <https://burckhardtsource.org/>

Gibson, James J., ‘The Theory of Affordances’, in *Perceiving, Acting, and Knowing. Towards an Ecological Psycology, ed. by Robert Shaw and John Bransford* (Hillsdale, NJ u.a, 1977), 67–82 <https://monoskop.org/images/c/c6/Gibson_James_J_1977_1979_The_Theory_of_Affordances.pdf>

Goody, Jack, ‘What’s in a List?’, in *The Domestication of the Savage Mind*, (Cambridge: Cambridge University Press, 1977), 74–111

Hendler, James, Deborah McGuiness, Richard Fikes, and Lynn Andrea Stein, ‘DAML-ONT: An Ontology Language for the Semantic Web’, in *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential,* ed. by Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, 65–93, 2005 <http://www.ksl.stanford.edu/people/dlm/papers/daml-ont-semantic-web.htm>

Knublauch, Holger, and Dimitris Kontokostas, ‘Shapes Constraint Language (SHACL)’, W3C Recommendation, 20 July 2017 <https://www.w3.org/TR/shacl/>

Huitfeldt, Claus, and C. Michael Sperberg-McQueen, 'TexMECS', 2001 <http://mlcd.blackmesatech.com/mlcd/2003/Papers/texmecs.htm>

'Hypothes.is', 2011–2020 <https://web.hypothes.is/>

Ide, Nancy, and Keith Suderman, 'The Linguistic Annotation Framework: A Standard for Annotation Interchange and Merging', in *Language Resources and Evaluation* 48,3 (2014), 395–418 <https://doi.org/10.1007/s10579-014-9268-1>

'ISO-LAF: Language resource management – Linguistic annotation framework (LAF). ISO 24612:2012' (ISO, 2012)

Ivanovs, Aleksandrs, and Aleksey Varfolomeyev, 'Some Approaches to the Semantic Publication of Charter Corpora. The Case of the Diplomatic Edition of Old Russian Charters', *Digital Diplomatics. The Computer as a Tool for the Diplomatist?*, ed. by Antonella Ambrosio, Sébastien Barret, and Georg Vogeler (Köln: Böhlau, 2014), 149–68

Kamzelak, Roland, 'Digitale Editionen im semantic web. Chancen und Grenzen von Normdaten, FRBR und RDF', in *'Ei, dem alten Herrn zoll' ich Achtung gern', FS Joachim Veit zum 60. Geburtstag*, ed. by Kristina Richts and Peter Stadler (München: Allitera Verlag 2016), 423–35

Kuczera, Andreas, 'Digital Editions beyond XML – Graph-Based Digital Editions', in *HistoInformatics 2016. Poceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016)*, ed. by M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van Den Bosch, CEUR Workshop Proceedings 1632 (Budapest: CEU 2016a), 37–46

———, 'Graphbasierte digitale Editionen', *Mittelalter: Opuscula* (blog), 19. August 2016b <https://mittelalter.hypotheses.org/7994>

Lacy, Lee W., *OWL: Representing Information Using the Web Ontology Language* (Victoria, BC: Trafford Publishing, 2005)

Langefors, Börje, *Theoretical Analysis of Information Systems* (Lund, 1966)

Morbidoni, Christian, and Alessio Piccioli, 'Curating a Document Collection via Crowdsourcing with Pundit 2.0', in *The Semantic Web: ESWC 2015 Satellite Events*, ed. by Fabien Gandon, Christophe Guéret, Serena Villata, John Breslin, Catherine Faron-Zucker, and Antoine Zimmermann, Lecture Notes in Computer Science 9341 (Cham: Springer International Publishing, 2015), 102–6 <https://doi.org/10.1007/978-3-319-25639-9_20>

Neill, Ian, 'The Codex. An Atlas of History', the-codex.net, 2013–2019 <http://web.archive.org/web/20190506014532/http:/the-codex.net/>

Neo4j, 2010–2020 <https://neo4j.com/>

Net7, *Pundit - Pin the Web*, 2015–2020 <http://thepund.it/>

Neumann, Gerald, Fabian Körner, Torsten Roeder, and Niels-Oliver Walkowski, 'Personendaten-Repositorium', *Berlin-Brandenburgische Akademie der Wissenschaften. Jahrbuch* 2010 (2011), 320–26

Nguyen, Vinh, Olivier Bodenreider, and Amit Sheth, 'Don't Like RDF Reification? Making Statements about Statements Using Singleton Property', in *Proceedings of the International World-Wide Web Conference. International WWW Conference*, 2014, 759–70 <https://doi.org/10.1145/2566486.2567973>

Nguyen, Vinh, Olivier Bodenreider, Krishnaprasad Thirunarayan, Gang Fu, Evan Bolton, Núria Queralt Rosinach, Laura I. Furlong, Michel Dumontier, and Amit Sheth, 'On Reasoning with RDF Statements about Statements using Singleton Property Triples', *arXiv:1509.04513 [cs]*, 2015 <http://arxiv.org/abs/1509.04513>

NIE-INE, 'Machine Interpretable and Interoperable Semantics for Humanities', e-editiones, 2020 <http://e-editiones.ch/>

Norman, Donald A., *The Design of Everyday Things* (New York: Doubleday, 1988)

Odgen, Charles K., and Ivor Armstrong Richards, *The Meaning Of Meaning : A Study of the Influence of Language upon Thought and of the Science of Symbolism* (London and New York, 1923)

Pasin, Michele, and John Bradley, 'Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach', *Literary and Linguistic Computing* 30,1 (1. April 2015), 86–97 <https://doi.org/10.1093/llc/fqt037>

Pelagios Network, Exeter University, Humboldt Institute for Internet and Society, The Open University, and University of London, 'Recogito', Software. Recogito, 2014–2020 <https://recogito.pelagios.org/>

Peroni, Silvio, *LODE - Live Owl Documentation Environment.* Webservice, 2012–2020 <https://essepuntato.it/lode/>

Petris, Marco, Jan Christoph Meister, Jan Horstmann, Janina Jacke, Christian Bruck, Mareike Schumacher, and others, *CATMA. Computer Assisted Text Markup and Analysis*, 2008–2020 <https://catma.de/>

Pichler, Alois and Zöllner-Weber, Amélie, 'Sharing and Debating Wittgenstein by Using an Ontology', *Literary and Linguistic Computing*, 28.4 (2013), 700–707 <https://doi.org/10.1093/llc/fqt049>

Pichler, Alois (ed.), *Wittgenstein Ontology*, The Wittgenstein Archives at the University of Bergen (WAB), 2020 <http://wab.uib.no/cost-a32_philospace/wittgenstein.owl>

Pollin, Christopher, and Georg Vogeler, 'Semantically Enriched Historical Data : Drawing on the Example of the Digital Edition of the ‚Urfehdebucher der Stadt Basel'', 2014, in *WHiSe 2017. Workshop on Humanities in the Semantic Web, 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*, ed. by Alessandro Adamou, Enrico Daga, and Leif Isaksen, CEUR Workshop Proceedings 2014. (Budapest: CEUR, 2017), 27–32

Poupeau, Gautier, 'De l'index nominum à l'ontologie. Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques?', in *Digital Humanities 2006. The First ADHO International Conference: Conference Abstracts. Université Paris-Sorbonne Paris 2006* (s.l.: ADHO , 2006), 161–64

Quillian, M. Ross, 'Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities', *Behavioral Science* 12,5 (1967), 410–30 <https://doi.org/10.1002/bs.3830120511>

Renear, Allen H, Steven J. DeRose, David G. Durand, and Elli Mylonas, 'What is text, really?', *Journal of Computing in Higher Education* 1,2 (1990), 3–26

Rennau, Hans-Jürgen, 'RDFe – expression-based mapping of XML documents to RDF triples', in *XML Prague 2019. Conference Proceedings*, Prag 2019, 381–404 <http://archive.xmlprague.cz/2019/files/xmlprague-2019-proceedings.pdf#page=393>

Burghartz, Susanna, Jonas Sagelsdorff, and Sonia Calvi, 'Jahrrechnung Stadt Basel 1545/1546', in *Jahrrechnungen der Stadt Basel 1535 bis 1610 – digital*, ed. by Susanna Burghartz, 2015. <http://gams.uni-graz.at/o:srbas.1545>

Sahle, Patrick, *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung*. Schriften des Instituts für Dokumentologie und Editorik 9 (Norderstedt: Books on Demand, 2013)

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young, 'Web Annotation Data Model', W3C Recommendation, 23 February 2017 <https://www.w3.org/TR/annotation-model/>

Schmidt, Desmond, and Domenico Fiormonte, 'A Fresh Approach to Textual Variation', *Digital Humanities 2006. Conference Abstracts* (s.l.: ADHO, 2006), 193–96

Schmidt, Desmond, 'The Inadequacy of Embedded Markup for Cultural Heritage Texts', *Literary and Linguistic Computing* 25,3 (2010), 37–56 <https://doi.org/10.1093/llc/fqq007>

Schmidt, Desmond, and Robert Colomb, 'A data structure for representing multi-version texts online', *International Journal of Human-Computer Studies* 67,6 (1 June 2009), 497–514 <https://doi.org/10.1016/j.ijhcs.2009.02.001>

Schrade, Torsten, *digicademy/xtriples* (Version 1.4.0). Zenodo, 2019 <http://doi.org/10.5281/zenodo.2604986>

simogih.org. *Le projet symogih.org : un système modulaire de gestion de l'information historique. Platform to store primary data*, 2012–2020 <http://symogih.org/>

Sperberg-McQueen, C. Michael, and Claus Huitfeldt, 'GODDAG: A Data Structure for Overlapping Hierarchies', in *Digital Documents: Systems and Principles, 8th International Conference on Digital on Documents and Electronic Publishing, DDEP 2000, 5th International Workshop on the Principles of Digital Document Processing, PODDP 2000, Munich, Germany, September 13–15, 2000, Revised Papers*, ed. P. King and E.V. Munson, Lecture Notes in Computer Science 2023 (Berlin: Springer, 2000), 139–60 <https://doi.org/10.1007/978–3-540–39916–2_12>

TCF Weblicht, 'The TCF Format', WebLichtWiki, 12 February 2015 <https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format>

TEI Consortium, '<witness>', P5: Guidelines for Electronic Text Encoding and Interchange. TEI <Text Encoding Initiative>, 2020a <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-witness.html>

———, 'Non-Hierarchical Structures', P5: Guidelines for Electronic Text Encoding and Interchange, 2020b <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html>

———, 'TEI P5: Guidelines for Electronic Text Encoding and Interchange', 13 February 2020 <https://web.archive.org/web/20200508073512/https:/www.tei-c.org/release/doc/tei-p5-doc/en/html/>

Tennison, Jeni, 'Overlap, Containment and Dominance', 6 December 2008 <https://www.jenitennison.com/2008/12/06/overlap-containment-and-dominance.html>

Tennison, Jeni, and Wendell Piez, 'The Layered Markup and Annotation Language (LMNL)', Montreal, Canada, 2002 <http://xml.coverpages.org/LMNL-Abstract.html>

Thaller, Manfred, 'kleio', kleio - on the web, 2003–2009 <http://web.archive.org/web/20090509075526/http:/www.hki.uni-koeln.de:80/kleio/>

———, 'What is a text within the Digital Humanities, or some of them, at least?', in *DH2012* -

Conference Abstracts. Hamburg, 2012 <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/beyond-embedded-markup>

Thompson, Henry, and Dave McKelvie, 'Hyperlink se-mantics for standoff markup of read-only documents', in *Proceedings of SGML Europe'97* (1997) <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>

Vallance, Edward, 'The 1723 Oath Rolls in England: An Electronic Finding List', *History Working Papers Project* (2013) <http://www.historyworkingpapers.org/?page_id=373>

Van Zundert, Joris, and Tara L. Andrews, 'Apparatus vs. Graph: New Models and Interfaces for Text', in *Interface Critique*, ed. by F. Hadler and J. Haupt, Kaleidogramme 139 (Berlin: Kulturverlag Kadmos, 2016), 183–205

Verborgh, Ruben, and Jos De Roo, 'Drawing Conclusions from Linked Data on the Web: The EYE Reasoner', *IEEE Software* 32,3 (2015), 23–27 <https://doi.org/10.1109/MS.2015.63>

Vogeler, Georg, 'The 'Assertive Edition', *International Journal of Digital Humanities* 1,2 (1. July 2019), 309–22 <https://doi.org/10.1007/s42803–019–00025–5>

Vogeler, Georg, and Christopher Pollin, 'DEPCHA - Alpha Version', DEPCHA - Digital Edition Publishing Cooperative for Historical Accounts. Alpha-Version, 2018–2020 <http://gams.uni-graz.at/context:depcha>

W3C, 'SKOS Simple Knowledge Organization System Reference. W3C Recommendation', Herausgegeben von Alistair Miles und Sean Bechhofer, 18th August 2009 <https://www.w3.org/TR/2009/REC-skos-reference-20090818/>

W3C OWL Working Group, 'OWL 2 Web Ontology Language Document Overview (Second Edition)', W3C Recommendation, 2012 <https://www.w3.org/TR/owl-overview/>

W3C, 'RDF 1.1 Primer. W3C Working Group Note', W3C Working Group Note, 2014 <https://www.w3.org/TR/rdf11-primer/#section-blank-node>

Wettlaufer, Jörg, 'Der nächste Schritt? : Semantic Web und digitale Editionen', in *Digitale Metamorphose: Digital Humanities und Editionswissenschaft*, ed. by Roland S. Kamzelak and Timo Steyer, Zeitschrift für digitale Geisteswissenschaften Sonderband 2, 2018 <http://zfdg.de/sb002_007>

Zeldes, Amir, Florian Zipser, and Arne Neumann, 'PAULA XML Documentation - Format Version 1.1', 21 January 2013 <http://www.sfb632.uni-potsdam.de/paula.html>

Zeller, Hans, 'Befund und Deutung. Interpretation und Dokumentation als Ziel und Methode der Edition', *Texte und Varianten. Probleme ihrer Edition Und Interpretation*, ed. by Gunter Martens and Hans Zeller (München: Beck, 1971), 45–90

Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, *GAMS - Geisteswissenschaftliches Asset Management System. Publication plattform*, 2014–2020 <http://gams.uni-graz.at/>

Formal Models

Formal Semantics for Scholarly Editions

Hans Cools (†), Roberta Padlina

Abstract

In the NIE-INE project based at the University of Basel, Switzerland, a national IT-infrastructure is being developed to support scholarly editions. In a first phase, it focuses on data representation and publication, and in a second phase on online editing. The paper deals with the Semantic Web technological (SWT) part, based on W3C standards, used to express edition project data and related knowledge in a formal way. A section on the project objectives discusses the benefits of formalizing the editions, ensuring the FAIR-principles. The hurdles of this process are also described. The state of the art deals with the components of the infrastructure. The development has a complex group of dependencies, of which the core technology provider (graph database and API) has the most impact. A first step, and major part of the SWT, is the development of formal vocabularies (ontologies) to express the editions' semantics. Also, external generic ontologies (e.g., CIDOC-CRM) are used, together with basic modeling patterns. The important features of the modelling process are described, for example, that expressing data in machine-interpretable standard languages requires explicitness. The results describe the differentiation and integration of ontologies, which are highly reusable for future projects. Different ontological graphics are used for different purposes.

1 Introduction

The aim of this paper is to present the implementation and development of Semantic Web technology (SWT) within the ongoing project *National Infrastructure for Editions* (NIE-INE 2019a).

The overall goal of NIE-INE, funded by swissuniversities[1] in 2016, is to create a Swiss infrastructure that includes an environment for the online publishing and editing of scholarly editions produced by projects in the Humanities. The chosen core technologies are the SWT standards developed by W3C (W3C 2001). The work comprises creating vocabularies or ontologies based on these standards, to convert

[1] https://www.swissuniversities.ch/en/themen/digitalisierung/p-5-wissenschaftliche-information/projekte/nie-ine.

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 97–124.

existing or new scientific data models and research data into formal machine interpretable expressions for long-term preservation, as well as further processing with machine reasoning.

Semantic Web technologies are particularly suitable for representing highly complex entities (like scholarly editions) and their relationships. Since questions of meaning and interpretation are central in the Humanities, it is essential to provide scholars with tools to express their research data and results in a formal, explicit, and self-descriptive way – and this is exactly where SWT proves its efficacy. The implementation of this technology initially corresponds to the creation of ontologies which, in the Semantic Web context, are formal data models that specify the concepts (classes), and the relations between concepts (properties), belonging to a domain knowledge. The W3C gives two concise definitions of an ontology in SWT: "a conceptualization of a domain to enable knowledge sharing" (W3C 2009) and "a representation of terms and their interrelationships" (W3C 2004b).

The most innovative aspect of the Semantic Web, in comparison to traditional formats and standards (e.g., TEI[2]/XML[3]), is that in an ontology, not only entities and their relations (linked data) are formally expressed, but also the tacit domain knowledge that forms the foundation of research data (for example, editorial principles and decisions, domain assumptions and perspectives, methodologies and workflows, and so on). This knowledge is crucial for an overall understanding, but is often implicitly hidden in the data. Texts and their editions, for example, are embedded in complex webs of discourse and narratives setting multidimensional relations (Robinson 2013, 107; Gabler 2010, 44). There is a well-known tension between two editorial perspectives (text-as-work and text-as-document), but, even if these perspectives set different interpretation contexts, they are indissolubly connected, and a scholarly edition must account for both (Robinson 2013, 123). SWT is precisely designed to cope with a wide range of information variance, and to allow multiple (even opposite) viewpoints to coexist within a single system. The conceptual description of the document-text-work triad, as well as concepts like authority, intention, agency and meaning, can hugely vary from one edition to another, but, from the moment that these fundamentally interpretative conceptualizations are explicitly formalized, the Semantic Web allows them to exist in a common semantic space wherein they are also explicitly related, and not locked in self-contained models or semantic silos.

Of course, to introduce variance, you first need commonality, and this is why knowledge formalization requires a minimal commitment and consensus on the part of domain experts. Achieving this is not without challenges, but the effort required

[2] The Text Encoding Initiative is a consortium that maintains a standard for the representation of text in digital form, http://www.tei-c.org/index.xml.

[3] The eXtensible Markup Language is another standard maintained by the W3C, https://www.w3.org/TR/2008/REC-xml-20081126/.

pays off, as it dramatically increases the expressiveness and the reusability of data and domain knowledge. It also eases interdisciplinary exchange, representing a clear benefit for other existing or future editions and humanities projects. Indeed, the lack of consistent authority files is one of the major challenges that Linked Open Data is facing at the moment. Therefore, the adoption of SWT will facilitate the establishment of formal linked data encompassing the FAIR-principles[4]. We believe that developing and implementing SWT represents a service for research in the Humanities, which, given the ever-increasing amount of digital data that scholars are producing, will prove more and more crucial.

2 Objectives

The main objective of the NIE-INE project in using Semantic Web technologies is the formal expression of scholarly editions, enabling semantic interoperability across editing projects, and, in the long run, creating a semantic space for an interdisciplinary formal knowledge domain for the Humanities. This interoperability is meant for both humans and machines, and, hopefully, will contribute to the long-term preservation of scientific data. To the best of our knowledge, the development of a variety of domain-specific ontologies for scientific editing in Humanities is quite a new niche, in which no comparable work has been done. The aim of NIE-INE is to create such a new digital environment, via the synthesis of existing technologies, and by creating new components i.e., formal ontologies, queries and rules, as well as a back- and front-end environment.

NIE-INE adheres to the FAIR-principles for scientific data management: scientific data should be **f**indable, **a**ccessible, **i**nteroperable, and **r**eusable. This requires appropriate data storage, editing and publication of scientific data.

The advantages of SWT translate to the FAIR-principles as follows:

- *Accessibility* is a feature (not equal to openness) of the Internet and Web technology, thus also of SWT.
- On the Internet and Web, resource locations are *findable* by their identifiers (URL[5]). On the Semantic Web all resources are identified (IRI[6]). Having an identifier ensures the ability to talk about the same thing, and having ontological elements as semantic identifiers enables mutual understanding.
- *Reusability* is enabled by the use of explicit, self-descriptive semantics to express information, with minimal ambiguity, in formal ontologies, providing data transparency and quality assurance. The terminological and conceptual consensus

[4] The FAIR-principles are: Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016).

[5] Uniform Resource Locator.

[6] Internationalized Resource Identifier. A URL is a type of IRI.

process for including domain knowledge in the ontologies also contributes to this principle.

- Due to the built-in logic (Berners-Lee 1998) of the formal standard languages, the ontology and data expressions are machine interpretable, enabling semi-automated *semantic interoperability*, going beyond the mere linking of hardware and software. The abstract expressions are natural language independent, enabling global information exchange. Thus, it is possible to link data from disparate sources and knowledge domains, adding domain knowledge and, therefore, facilitating reuse and multidisciplinarity.

The machine-interpretable semantics of the formal ontologies, data, and rules hugely increases the information expressiveness, and also allows rule-based machine reasoning. This kind of Artificial Intelligence further improves data quality through consistency and validity constraints checks. It is used to enhance data expressiveness by means of many kinds of calculations, for example, interval calculus for temporal reasoning, calendar conversion, data comparison and analysis.

The formal data models in SWT are much more flexible, and independent from the technological implementation, than the models in relational databases or XML; new data models can therefore be implemented in a more sustainable way. In comparison to SQL[7] databases, where column names and identifiers are local, IRIs make links and values explicit and globally identifiable. Another main advantage of the Semantic Web model is that it is non-hierarchical. That is why it can represent complex entities and their relationships in a better way than is possible in a hierarchical, tree-based data model like XML. Thus, SWT can overcome problems resulting from XML implementations (e.g., developing a native stand-off solution), which can also have repercussions for the usage of the TEI guidelines. TEI is considered the standard for the representation of textual resources, but, being an XML-based language, TEI is more a syntactic and serialization language than a formal semantic model. TEI's usage in the community is largely influenced by circumstances and pragmatic factors, since the TEI guidelines allow for many ways of expressing the same textual feature, while the same TEI element can be subject to different interpretations depending on the context. Moreover, in XML and TEI the text has to be organized according to a single "ordered hierarchy of content objects" (OHCO), which makes it very difficult to represent concurrent hierarchies and overlapping features of textuality (DeRose et al. 1990; Renear et al. 1996). This is why some proposals to formalize the semantics of the TEI framework have already been made (Ciotti 2018).

The implementation of SWT, including the migration of existing data to the RDF format, isn't effortless, but entails some hurdles and challenges, of which it is important to be aware:

[7] Structured Query Language, used for managing data in relational databases.

- SWT is often perceived as replacing existing database systems (mainly SQL and XML). However, being one of the most modern data technologies, it is, rather, offering a complementary environment to the classical databases and data processing tools.
- SWT is not yet perceived as a standard in the Humanities. On one hand, being situated in the back-end as a middleware, it is not directly visible for non-IT specialists; on the other hand, being declarative and quite different from other, imperative programming, it is perceived as an extra challenge: notably the "open world assumption" often makes people feel uncomfortable.
- SWT needs the implementation of a complex environment with different elements covering it end-to-end, from input database, through formal semantic data model, to user interface. It is not merely another data model, but also includes reflections on semiotics, semantics, linguistics (in relation to different natural languages), logic, and IT, comprising standard formal languages and their ontologies, a query language, and rule-based machine reasoning. According, a learning curve has to be taken into account.
- Developing an ontology for a new given domain also requires an initial overhead building up the foundation of a semantic space, needing discussions and consensus with domain experts.

3 State of the Art

In the NIE-INE project, different technological domains are brought together: SWT, an overall Web framework called *inseri* (former NIE-OS)[8], as well as the coordination and development of digital editing tools[9]. For development using SWT, the W3C formal standard languages are used:

- the Resource Description Framework (RDF) (W3C 2004a; W3C 2014a), RDF Schema (RDFS) (W3C 2014b), and Web Ontology Language (OWL) (W3C 2004b; W3C 2012), are used to express formal ontologies and data;
- the RDF Query Language SPARQL (W3C 2013), for information retrieval from an RDF graph database;
- Notation3 (N3) (W3C 2011), to express formal inference rules for machine reasoning.

These standards are based on set theory, model or interpretation theory, and first order logic.

[8] The source code is published in open access as a Git repository, https://github.com/nie-ine/inseri. A test instance is available at http://test-nieos.nie-ine.ch/.

[9] The presentation of the state of the art of *inseri* and editing tools is beyond the scope of this paper.

The machine reasoner used is Euler Yet another proof Engine (EYE) (Verborg & De Roo 2015).

The NIE-ontologies are serialized in Turtle (W3C 2014c), a subset of N3, which is much more human-readable than the initial RDF-XML standard format.

Of course, the NIE-INE project did not start creating formal ontologies from scratch. Concerning modelling OWL-ontologies in general, we refer to Allemang and Hendler (2011), and in Humanities we refer to Eide and Ore (2018).

Standard ontologies such as Friend of a Friend (FOAF) (Brickley & Miller 2014), Dublin Core Metadata Initiative (DCMI 1995), and Simple Knowledge Organization System (SKOS, Miles & Bechhofer 2009) are used, together with some of the basic ontologies of the SWEET series (ESIP 2019).

Concerning domain-oriented terminology for the Humanities, NIE-INE ontologies are based on:

- International Committee for Documentation – Conceptual Reference Model (CIDOC-CRM) (CIDOC 2006);
- Functional Requirements for Bibliographic Records object-oriented extension to the CIDOC-CRM (FRBRoo) (Dunsire 2014; IFLA 1998).

4 Methodologies

4.1 Dependencies

In the NIE-INE project there are a series of dependencies that directly influence the implementation of SWT, especially the modelling of ontologies:

- The project makes use of the W3C standards, the aforementioned standard and existing domain ontologies, basic modelling patterns, and best practices.
- NIE-INE supports overall eleven projects, eight of which are currently using SWT. Moreover, several new projects have already indicated their willingness to collaborate with us. The input dependency is represented by the original data model and data, mostly in XML (5 projects, which primarily, but not exclusively, use TEI), SQL (2 projects), or mixed (1 project).
- Last but not least, there is a dependency on the tacit knowledge of the domain specialists working on the editing projects, which is required for the implementation of domain knowledge beyond the restricted database models. To uncover this knowledge, a commitment to modelling on the part of domain specialists is needed, in order to reach consensus on semantics beyond their own specific research objectives.

All these dependencies represent a major challenge and, initially, a substantial overhead, but the return on investment (ROI) is big, and will be even bigger the more project database models are formalized using SWT. This formalization will also contribute to the more general cross-project semantics, removing the need to model the same concepts for new projects multiple times. Reaching a wide consensus (ideally, on an international scale) on domain terminology contributes to the ROI and semantic interoperability.

4.2 Preliminary Analysis

Knowledge of the W3C standards and the aforementioned standard and basic ontologies, together with knowledge of input data model formats like XML, especially TEI, and SQL, was present from the start of the NIE-INE project. Existing domain ontologies, like CIDOC and FRBRoo, and the applied framework[10] had to be analysed. For every edition project that NIE-INE supports, an in-depth investigation of the data model is also provided. This implies a deep understanding of digital scholarly editing, requiring models to be produced as the result of an iterative epistemological hermeneutic process (Pierazzo 2015).

4.3 Basic Modelling Patterns

By conceiving or adopting basic modelling patterns that will be abundantly used in ontologies, modelling begins before an ontology is actually declared. Such patterns comprise one or more basic concepts and their sets of relations (properties). Examples of such patterns are *event* and *role*[11]. *Event* and its consecutively derived concepts *process*, *action*, and *procedure* are considered as four-dimensional entities described in time (start, end) and place (geographic names and coordinates), having inputs and outputs (process), with agents (person, organization, and even software) having roles (functions) in an action. *Procedure* is a particularly useful pattern. It captures both the perspectives of a time-space entity and the result thereof, making it possible to clearly distinguish between, for example, an edition as a procedure (editing), and an edition as a resulting product; it is therefore well-suited for describing scientific editing as a variety of procedures, involving agents with different roles, and having a variety of in- and outputs.

Moreover, in the Humanities, it is essential that data are positioned in time and space. Therefore, indicators for dates and places are primordial. Even approximate indicators can be used (which is especially useful for data concerning ancient times).

A more domain-oriented modelling pattern concerns the concept of *reference*, which

[10] https://www.knora.org/

[11] For more information on basic modelling patterns, http://e-editiones.ch/ontology-modeling#basic-modeling-patterns.

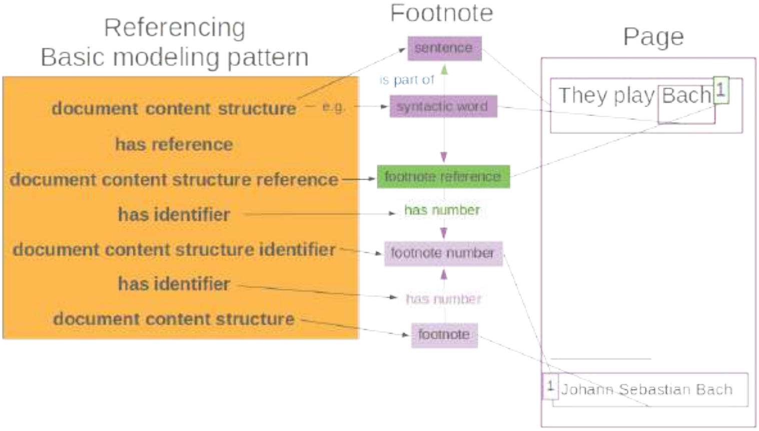


Figure 1. Basic modelling pattern for the concept *reference*.

is especially important for the Humanities. Figure 1 shows the concepts involved using the *footnote* entity as an example.

4.4 Development

Ontologies, SPARQL queries and N3 rules are created with a text editor, and ontologies are checked on syntax and logic consistency in the open source editor Protégé[12].

The whole development evolves in a very iterative way, requiring the connection of different roles and expertise. It is very important to have regular discussions between NIE-INE modelers (4) and edition projects contacts (5), as well as with domain specialists (12), as it is impossible to capture the required project semantics all at once.

4.5 Identification and Structure

The basic expression unit structure in SWT is the triple, consisting of a *subject*, a *property* (predicate), and an *object*, each being an IRI, or, in case of an object, also a plain or typed literal value. A set of triples makes a graph: an ontology and RDF data are graphs; the result of a SPARQL query is also a graph. Figure 2 shows an example of such an identification and serialization of a sentence in natural language, encoded in XML, and converted into RDF triples.

[12] Protégé is an ontology editor developed by the Stanford Center for Biomedical Informatics Research, https://protege.stanford.edu/.

Natural Language expression as literal:

```
''The person has the name Petrus Lombardus.''
```

In XML:

```
<person>
  <name>Petrus Lombardus</name>
</person>
```

RDF triples in Turtle:

```
example:personX a human:person . # 'a' is syntactic sugar for rdf:type
example:personX human:hasNameLiteral ''Petrus Lombardus'' .
```

Note:
`example:`, `rdf:` and `human:` are replacments of the respective namespace part of a full IRI, e.g. `<http://www.e-editiones.ch/ontology/human\#>`, making Turtle more redable.

Figure 2. Sentence in natural language and XML converted in RDF triples.

4.6 Modelling

Explicit Statements

Since the W3C OWL ontologies contain the built-in logic of the RDF model theory, they are machine interpretable. When basing an own ontology on these, its elements have to be declared in an explicit way, with minimal hidden assumptions, to keep them machine-interpretable. Being explicit also represents a good way to reveal flaws in the original or formal data model.

It is important to point out that this process of introducing standards and being explicit does not reduce the expressiveness of the data. On the contrary, RDF permits the co-existence of distinctive expressions of domain knowledge, as long as they are made explicit.

For human development and usage, ontologies and ontological elements obtain clear multilingual labels and a description. We think it is important to keep the latter concise, because the longer it is, the more scope there is for interpretation, rather than clarifying meaning.

Reification and Abstraction

CIDOC-CRM and FRBRoo come with extra levels of abstraction and reification. Examples of reification are the class `cidoc:E41_Appellation`, and the subclass `cidoc:E44_Place_Appellation`. An instance of the latter has an IRI – that is, it is a thing, not a string – which, in turn, has a name as a literal string value, e.g., "Lausanne". Similarly, all instances of the class `cidoc:E33_Linguistic_Object` have an IRI, which in turn are linked to literal expressions. Examples of abstraction levels are the basic classes `frbroo:F1_Work`, `frbroo:F2_Expression`, `frbroo:F4_Manifestation_Singleton`, and `frbroo:Item`.

Building further on these FRBRoo classes, in the NIE-INE ontologies there are e.g., four different classes dealing with *page*:

1. `information-carrier:Page`: a physical surface of a leaf, e.g., in a manuscript or a book;
2. `document:Page`: a content structure with information, e.g., text lines and graphs, as part of a document expression, on an information carrier page;
3. `text-structure:Page`: text structure as part of a document page and a text expression; and
4. `text-editing:Page`: text page of a scientific edition.

Middle-Out

Adhering to the principles *as simple as possible, as complex as needed,* and *the least ontologic commitment* (i.e., providing the necessary and sufficient semantics), the modelling faces a challenge in finding the right balance between ground elements and details. Deep grounding is provided by very basic or upper ontologies, e.g., CIDOC-CRM. On the other end of the spectrum, there is the rather ad hoc project-specific modelling in a stand-alone way, which is not useful for semantic interoperability. Middle-out modelling begins with some required points of depth and detail, in a way that makes it possible to easily extend ontologies.

Multitude of Namespaces

In creating a semantic space for broad applicability, ontology size, and consequently the number of ontologies developed, also matters. Having an *all-in-one* approach is rather difficult if the intended research-space is broad, implying multiple domains. A scientific project always needs general concepts such as *person*, *document*, *text*, *image*; these concepts are typically reused from more general ontologies. A project ontology can exist, but should be based on these general ontologies, and only contain project-specific elements, for example, the concept of *Dreissiger* (a set of about 30 verses) used in the *Parzival*-project.

Partitioning knowledge over ontologies upfront is also a challenge, but doing so is important in order to avoid the need for later division of an ontology. Moreover, new knowledge is constantly produced, possibly leading to deprecation of old knowledge and, thus, requiring the splitting of ontologies.

Property Oriented

Entities obtain meaning based on the way, and in the extent to which, they are related to other entities via properties. Databases often contain implicit, condensed, or shortcut semantics. In order to be explicit, it is important that these semantics are unravelled, and consequently that more properties are added: a simple example is the conversion of a name to a family name and a given name. Together with the previous modelling feature (multitude of namespaces), this leads to a *network* or a web of OWL ontologies and RDF data graphs.

Consensus

Last, and certainly not least, iterative discussions with the project domain specialists must lead to a minimal consensus about domain concepts. As previously stated, consensus is also essential for enabling semantic interoperability. This makes modelling a very *collaborative* and multidisciplinary activity.

4.7 Database to RDF Mapping

Once the semantics of a project are covered in the ontologies, the original data model can be mapped to RDF using the classes and properties defined in the ontologies. This is done in a spreadsheet, using the following methodology.

For XML, parent nodes are mapped to RDF classes, and child nodes and attributes are mapped to properties. Content and attribute values become property values of instances.

For SQL, table names are mapped to RDF classes, and column headers are mapped to properties. Keys are used to instantiate a row as a member of a class, with a series of properties mapping to the field values.

After the mapping is complete, a script is written (e.g., in XSLT or Python) to convert project data in XML or SQL to RDF, in order to import the data into an RDF database or triple store. During conversion, the data elements obtain an IRI, and are checked for consistency with the built-in logic of the applied W3C standards part.

Once in the triple store, RDF data can be checked using large SPARQL queries, covering the whole semantics of an edition. From this point forward, RDF data can be retrieved with specific SPARQL queries, converted to JSON and then either consumed by JavaScript-based web frontends, or further processed. Repeated representation or functionality can be supported by existing SPARQL query libraries, which can retrieve the necessary parts of RDF data.

The following is a summary of the tasks in the workflow of formal semantic modelling in NIE-INE:

1. Ontology development as servicexs and products
    - 1.1. Initiating new project
    - 1.2. General semantics emerged, but independent from individual edition project
    - 1.3. Modelling iteratively, multiple projects simultaneously
    - 1.4. Change management (e.g. dependency on application)
    - 1.5. Graphics
    - 1.6. Information
2. Mapping original data model to RDF using ontologies
    - 2.1. Test ontology XML-schema generation (application)
3. Import scripts for bulk data import
4. Test triple store data with SPARQL
5. Create SPARQL query sets for apps

Future:

6. Notation3 rules for machine reasoning
    - 6.1. Second setp (semantic conversion) in the 2-step formalization
7. 2-step formalization

The most demanding tasks are 1.2, 1.3, 2, 3, 6 and 7.

4.8 Future

Machine Reasoning

In the near future, we intend to apply machine reasoning on the formal data using the built-in logic of the formalisms. Besides the ontologies, these formalisms will comprise N3-rules for different purposes: to provide automated data quality assurance through consistency checks (e.g. cardinality of instance existence); to infer new data from existing data using the built-in domain knowledge of N3-rules, thus enhancing data expressiveness in a multidisciplinary environment; to perform temporal reasoning (interval calculus). Temporal reasoning can be applied, for example, in case of an event (basic modelling pattern) that is lacking start or end date, e.g., when a birthdate of a life (as an event) is missing. By defining a maximum life span, a certainty period can be established, calculating the earliest birthdate. Similarly, in a case where two different birthdates are mentioned for the same person, a certainty interval for the birth event, comprising both dates, can be calculated, together with the certainty interval of the derived life event, with the earliest birthdate as the start. Concerning time indicators NIE adopts the EDTF level 0 (Library of Congress 2019). The other

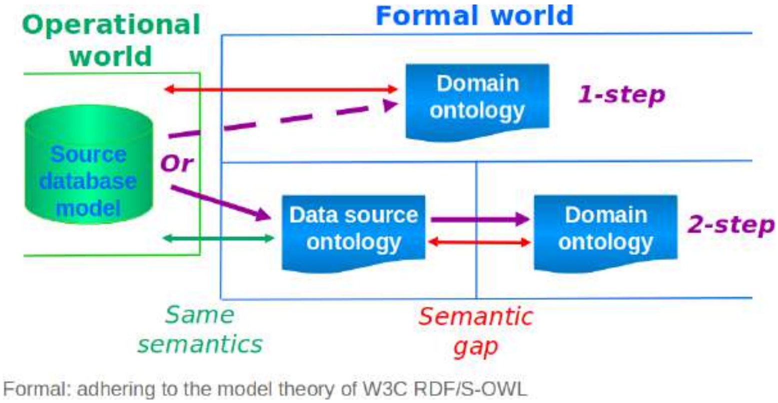


Figure 3. 1-step vs. 2-step formalization from source database model to formal domain ontology.

levels are already partially supported with N3-rules, and can be further implemented. RDF result sets of the machine reasoning process are then added to the triple store. As already mentioned, the reasoner being used is EYE[13].

2-Step Formalization

To formalize the data in a flexible and convenient way, a so-called 2-step formalization is planned for the near future, decoupling the database semantics from the formal domain knowledge. Until now, RDF data representing formal domain ontologies are created in one step, and have a direct link to the original source data. In a 2-step system, the first step corresponds to a 1-1 conversion of the original data to RDF, without interpretation, or making explicit all of the semantics (see Figure 3). The data can then be stored in this format in the RDF database. In a second step, the RDF data are converted to explicit, enriched semantics using domain ontologies and N3-rules. In this way, it is possible to enable semantic interoperability in a more decoupled way, that is less dependent on data source and application specificities.

Another example of temporal reasoning can be seen at the second step: the simple literal time indicators are converted to typed literal data values, i.e., *year*, *month*, and *day*, go from being three fields in a relational data table, to being represented as an interval of one year with a start and end `xsd:dateTime` (e.g., from *1886*, *9*, and *4* to start: `1886-09-04T00:00:00.0Z^^xsd:dateTime` and end: `1886-09-04T23:59:59.999997854Z^^xsd:dateTime`).

[13] EYE is an open source reasoning engine, https://github.com/josd/eye.

Number of → in ↓	Ontology	Rdfs:Class	owl:ObjectProperty owl:DatatypeProperty
Generic ontologies	37	671	497
Project ontologies	8	135	44
Total	45	806	541

Table 1. NIE ontologies in numbers.

Collaboration

It is also our intent to strengthen the collaboration with other projects using SWT, in particular with *histHub*[14], the Scholastic Commentaries and Text Archive (SCTA[15]), SARI (Swiss Art Research Infrastructure), and the Zentrum für Informationsmodellierung at the University of Graz[16].

5 Results

All *ontologies* created in the NIE-INE project are open source and published on GitHub as Turtle files (NIE-INE 2019b). There are two series: generic and project ontologies. The former is, *grosso modo,* further divided into four levels: 1) general domain, 2) general Humanities, 3) specific Humanities, and 4) external terminology and code systems ontologies (see Figure 4). They differ quite a lot in size, granularity and specificity. This division is somewhat arbitrary, meaning that it isn't formalized, but it is convenient to illustrate the articulation of ontologies and their interrelations. Table 1 shows the status of the modelling at the end of September 2019. The most populated ontologies are *human*, *information carrier*, *document*, *text*, *text-expression*, *text-structure*, *scholarly-editing*, *publishing*, *literature*, and *linguistic-morphology*. The *time*-ontology contains properties mainly for N3-rule declarations.

Ontologies can easily be added to or extended. In the development phase until end 2019, we still foresee possible splits of faster growing ontologies, and the transfer or replacement of elements.

Ontologies and their elements can be analysed from different perspectives and entry points. In NIE-INE, four kinds of graphics are used to illustrate the ontologies, to show classes and properties and instances in different ways.

[14] The aim of histHub is to establish and operate a research platform for the historical sciences, https://histhub.ch/.

[15] SCTA's aims is to connect and freely distribute the intellectual history of the scholastic tradition, https://scta.info/.

[16] https://informationsmodellierung.uni-graz.at/.

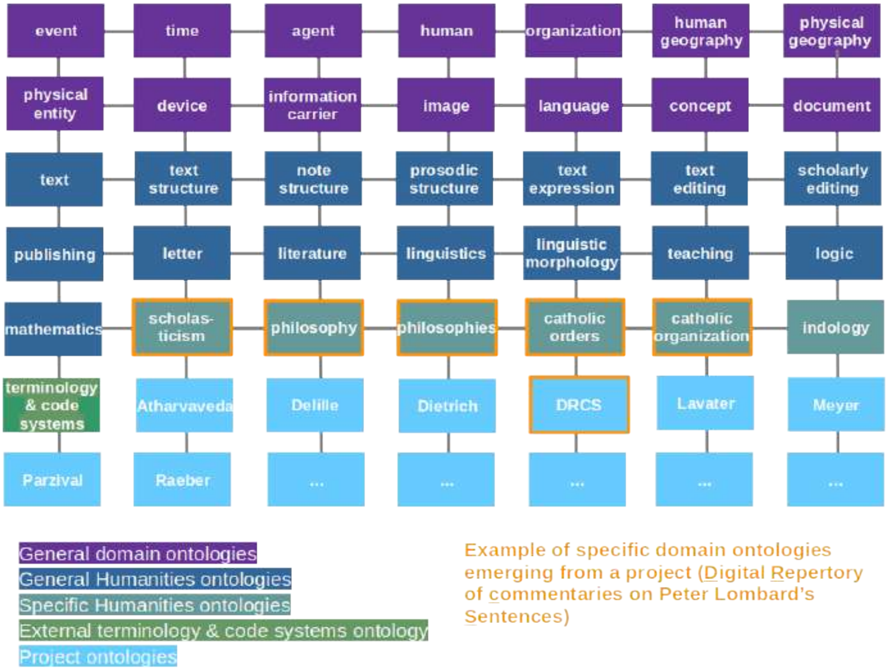


Figure 4. A simplified representation of the NIE-INE web of ontologies.

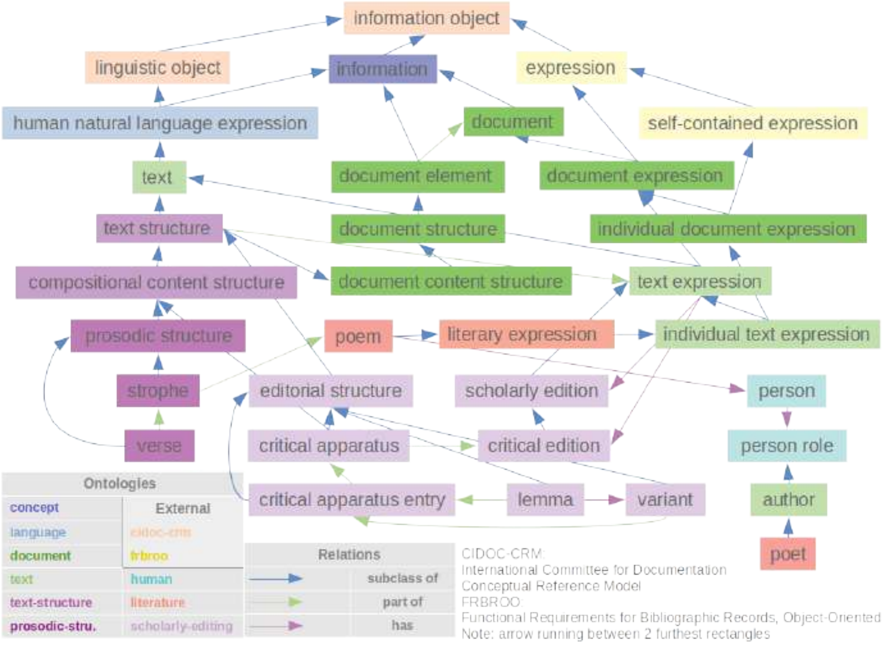
A first, manually created, graphic type (Figure 5) shows different core domain ontological elements in a simplified way, offering a broad overview of the dependencies between the ontologies, while enabling a focus on certain semantics, (e.g., on text and critical editing,) while also providing more general semantics.

A second type of graphic (Figure 6) centers one entity, e.g., *event*, relating it to elements from different domain ontologies, and adhering (in a reduced way) to the RDF structure, e.g., indicating prefixes and full property names. Classes are represented by circles; properties by rectangles. Classes from external ontologies are colored orange in the upper part. The subclass property is represented by a dotted arrow. This graphic is created with Grafo[17].

A third type of graphic (Figure 7) focuses on (a part of) an ontology, also representing properties as nodes. It is created with the open source SPARQL-visualizer[18],

[17] Available at https://gra.fo/.

[18] Available at https://github.com/MadsHolten/sparql-visualizer.



CIDOC-CRM:
International Committee for Documentation
Conceptual Reference Model
FRBROO:
Functional Requirements for Bibliographic Records, Object-Oriented
Note: arrow running between 2 furthest rectangles

Figure 5. Graphic representing core classes and properties from different domain ontologies.

filtering some ontological elements out (labels, comments, and blank nodes) to enhance readability. It is particularly suitable for discussions on terminology with domain specialists.

A fourth type (Figure 8), created with Protégé, depicts a subsumption tree of a merged group of ontologies. It is very useful to get a quick and broad hierarchical overview during the modelling stage. In the example, two trees are shown, building on CIDOC-CRM and FRBRoo. The left tree contains document expression, text expression, and subclasses. The right tree contains document structure, text structure, and subclasses (note: not all document structure subclasses are shown).

For further graphic examples we refer to the NIE-INE GitHub site. We will now discuss in more detail the different levels of ontologies.

5.1 General Domain Ontologies

Although all ontological classes are instantiable, the more general ones will often function as *glue* to search in an RDF-database with SPARQL queries, and to enhance

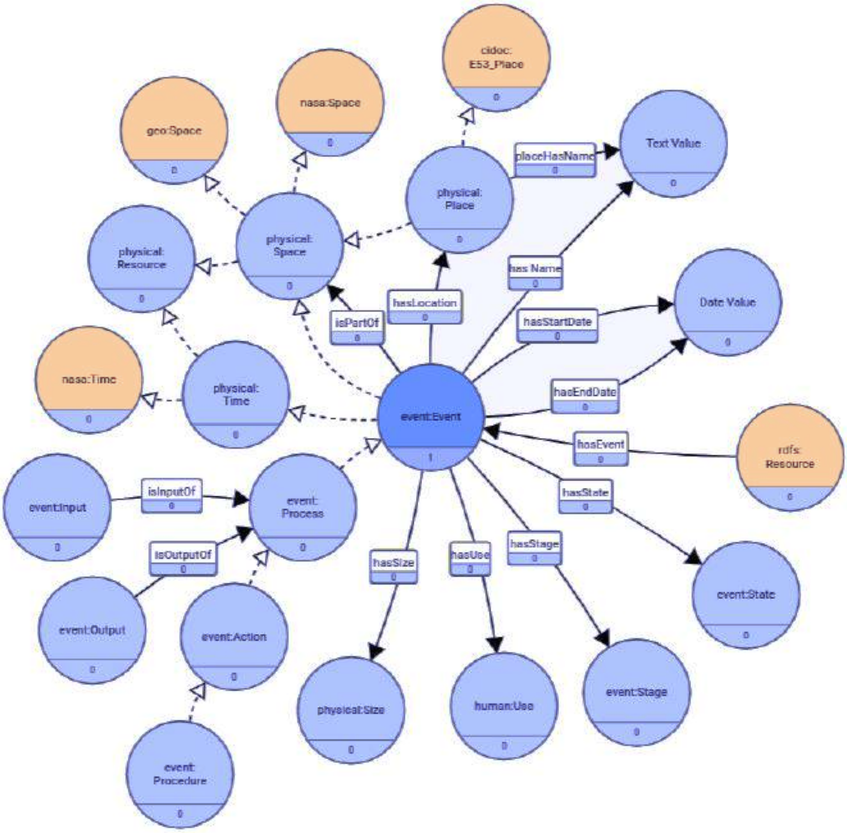


Figure 6. Graphic representing classes and properties from different ontologies concerning the concept *event*.

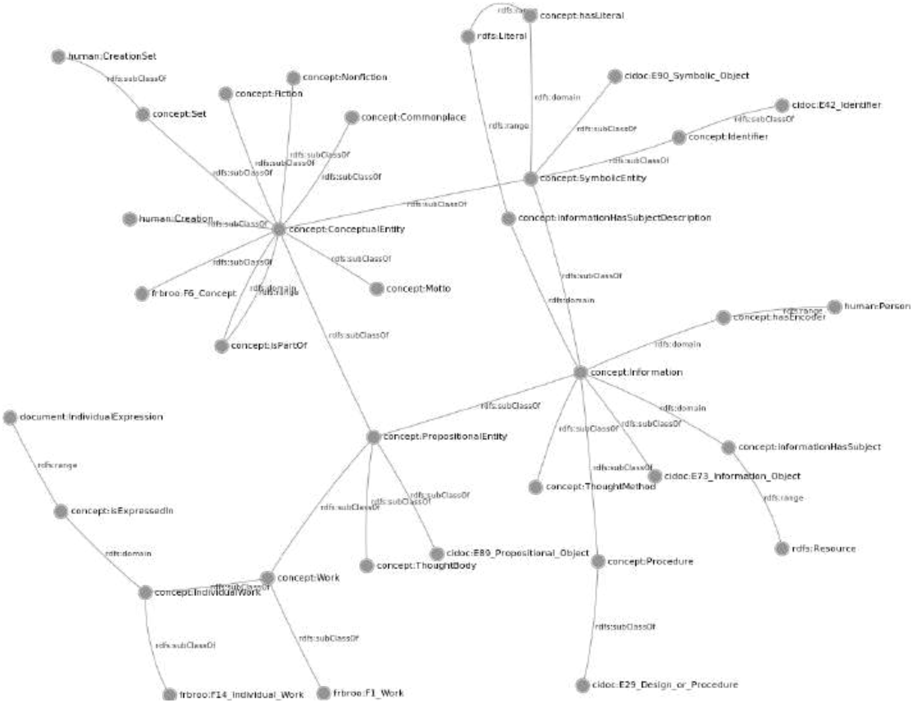
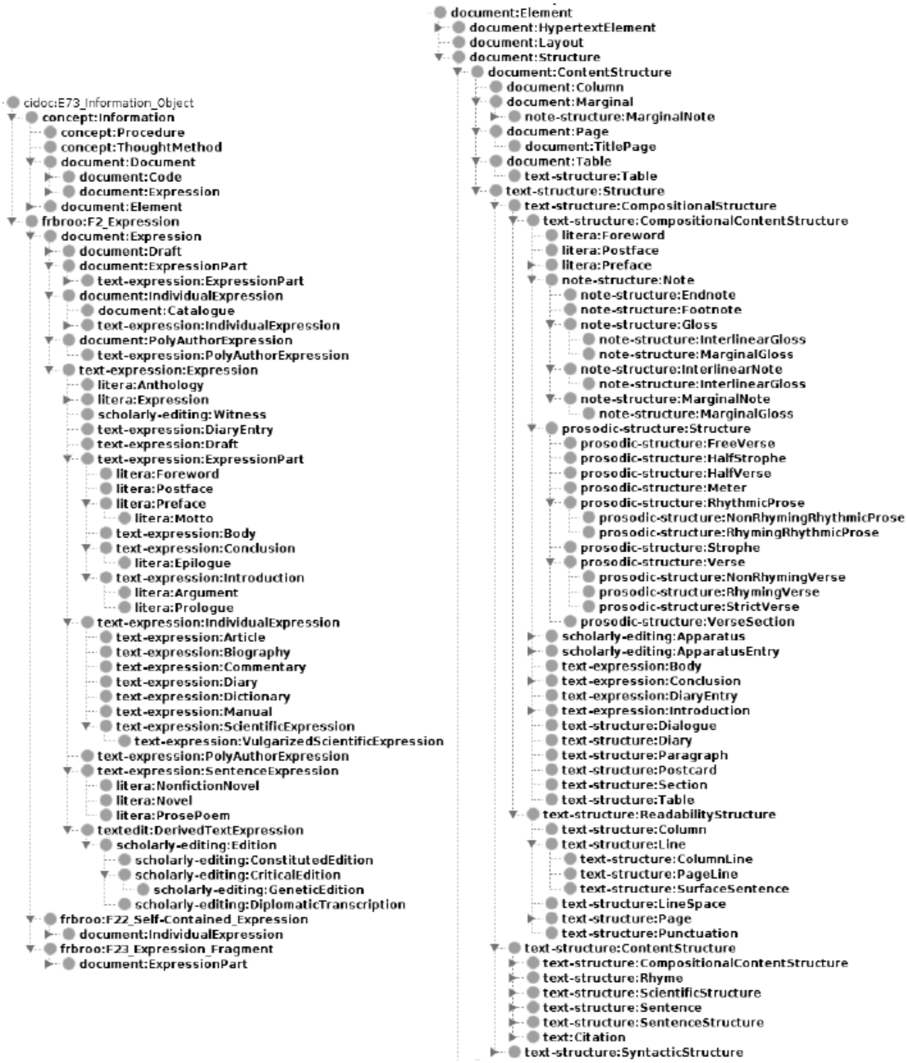


Figure 7. The concept-ontology.

machine reasoning, e.g., for subsumption (subclassing). For example, if a scholar wants to search for all the languages in which a text is translated, the property *is translated into* can be used without specifying the language. In other words, all instances of all subclasses of the class *human natural language* are retrieved. In another case, one would like to find all information carriers (e.g., manuscripts and prints) bearing creations from of a certain author, across more than one project: one project refers to prints existing in an archive and having a signature, and another project uses manuscripts preserved in a library with a manuscript identifier; in this instance, both the manuscripts and prints can be found with a super-property *is on carrier*.

As a general domain ontology, the *concept*-ontology (see Figure 8) describes, among other things, concepts created by a person as abstract ideas, e.g., symbolic and propositional, basing on CIDOC-CRM and FRBRoo. It contains entities such as *information*, *identifier*, and *thought-method*, and their relations to each other, and to other entities,



```
cidoc:E73_Information_Object
  concept:Information
    concept:Procedure
    concept:ThoughtMethod
    document:Document
      document:Code
      document:Expression
    document:Element
  frbroo:F2_Expression
    document:Expression
      document:Draft
      document:ExpressionPart
        text-expression:ExpressionPart
      document:IndividualExpression
        document:Catalogue
        text-expression:IndividualExpression
      document:PolyAuthorExpression
        text-expression:PolyAuthorExpression
      text-expression:Expression
        litera:Anthology
        litera:Expression
        scholarly-editing:Witness
        text-expression:DiaryEntry
        text-expression:Draft
        text-expression:ExpressionPart
          litera:Foreword
          litera:Postface
          litera:Preface
            litera:Motto
          text-expression:Body
          text-expression:Conclusion
            litera:Epilogue
          text-expression:Introduction
            litera:Argument
            litera:Prologue
        text-expression:IndividualExpression
          text-expression:Article
          text-expression:Biography
          text-expression:Commentary
          text-expression:Diary
          text-expression:Dictionary
          text-expression:Manual
          text-expression:ScientificExpression
            text-expression:VulgarizedScientificExpression
        text-expression:PolyAuthorExpression
        text-expression:SentenceExpression
          litera:NonfictionNovel
          litera:Novel
          litera:ProsePoem
        textedit:DerivedTextExpression
          scholarly-editing:Edition
            scholarly-editing:ConstitutedEdition
            scholarly-editing:CriticalEdition
              scholarly-editing:GeneticEdition
            scholarly-editing:DiplomaticTranscription
  frbroo:F22_Self-Contained_Expression
    document:IndividualExpression
  frbroo:F23_Expression_Fragment
    document:ExpressionPart
```

```
document:Element
  document:HypertextElement
  document:Layout
  document:Structure
    document:ContentStructure
      document:Column
      document:Marginal
        note-structure:MarginalNote
      document:Page
        document:TitlePage
      document:Table
        text-structure:Table
      text-structure:Structure
        text-structure:CompositionalStructure
          text-structure:CompositionalContentStructure
            litera:Foreword
            litera:Postface
            litera:Preface
            note-structure:Note
              note-structure:Endnote
              note-structure:Footnote
              note-structure:Gloss
                note-structure:InterlinearGloss
                note-structure:MarginalGloss
              note-structure:InterlinearNote
                note-structure:InterlinearGloss
              note-structure:MarginalNote
                note-structure:MarginalGloss
            prosodic-structure:Structure
              prosodic-structure:FreeVerse
              prosodic-structure:HalfStrophe
              prosodic-structure:HalfVerse
              prosodic-structure:Meter
              prosodic-structure:RhythmicProse
                prosodic-structure:NonRhymingRhythmicProse
                prosodic-structure:RhymingRhythmicProse
              prosodic-structure:Strophe
              prosodic-structure:Verse
                prosodic-structure:NonRhymingVerse
                prosodic-structure:RhymingVerse
                prosodic-structure:StrictVerse
              prosodic-structure:VerseSection
            scholarly-editing:Apparatus
            scholarly-editing:ApparatusEntry
            text-expression:Body
            text-expression:Conclusion
            text-expression:DiaryEntry
            text-expression:Introduction
            text-structure:Dialogue
            text-structure:Diary
            text-structure:Paragraph
            text-structure:Postcard
            text-structure:Section
            text-structure:Table
          text-structure:ReadabilityStructure
            text-structure:Column
            text-structure:Line
              text-structure:ColumnLine
              text-structure:PageLine
              text-structure:SurfaceSentence
            text-structure:LineSpace
            text-structure:Page
            text-structure:Punctuation
        text-structure:ContentStructure
          text-structure:CompositionalContentStructure
          text-structure:Rhyme
          text-structure:ScientificStructure
          text-structure:Sentence
          text-structure:SentenceStructure
          text:Citation
        text-structure:SyntacticStructure
```

Figure 8. Two subsumption trees representing classes from different ontologies concerning the concepts document expression and text expression on the left and document structure and text structure on the right.

e.g., persons. The *document*-ontology (see also Figures 5 and 8) describes documents and document structures, e.g., tables, identifiers, and references, such as footnotes. It also contains the abstract document expression as based on FRBRoo[19], and the different relations between document structures.

5.2 General Humanities Ontologies

The *general Humanities ontologies* comprise the core concepts of scholarly editions, and the major part of entities common to the individual models of single editions. The following is a description of five core vocabularies used for scholarly editing in Humanities (see Figures 5 and 8, above).

- Text: this ontology describes text as a human natural language expression, serialized in writeable form. It contains all kinds of text forms (e.g., written, typewritten, transcribed, and printed) and the roles of persons processing text (e.g., editor, annotator, and citer). It serves as the basis for all text-related ontologies: text-expression, text-structure, text-editing, and literature ontology.
- Text expression: the eponymous core concept bases (via the document-ontology) on FRBRoo, that abstracts text from its carrier (see Figure 8). The ontology defines further related roles (e.g., author and commentator) and general expression types (e.g., draft and commentary). It provides the basis for the literature- and scholarly-editing-ontology.
- Text structure: this ontology describes all kinds of textual structures (see Figure 8), e.g., syntactic, compositional, content, scientific, readability. They form an upper layer to enable more flexible and extensive search, as well as machine reasoning. More specific entities are word, sentence, paragraph, section, line, and text column. Extensions of this ontology are the prosodic-structure-ontology and the note-structure-ontology, containing entities such as verse and strophe, and marginal note and gloss, respectively. An important relation between structures is *part of*, which, by its transitive nature, enables searching and machine-reasoning in a way that does not require explicitly stating all possible relations between structures in the data, since they can be inferred via transitivity.
- Text editing and scholarly editing both of these terms describe the necessary semantics for editing, the latter extending the former with specific scholarly entities, e.g., diplomatic transcription, critical edition, different types of apparatus, lemma, variant, editorial comment, witness, siglum and so on. Also, related roles are declared, e.g., editor, glossator, and critical text editor. An extensive set of properties relates these entities to one other, as well as to text, text structure and expression elements (see Figures 5 and 8).

[19] FRBRoo provides the concept of expression, abstracted from its carrier.

- Publishing: this ontology describes classes and properties related to publishing, that is, to publication and its subclasses: printed and web publication, serial-like newspaper, periodical, magazine, etc. There is a substantial set of properties relating expressions and other entities to elements in this ontology.
- Literature: this ontology describes literary genres such as narrative, different kinds of poetry, and further different types of literary expressions (e.g., poem, hymn, novel) and their subclasses. It also contains related roles, e.g., poet and novelist. Different types of literary structures are declared, e.g., foreword, preface, prologue, and epilogue, and related properties (see Figures 5 and 8).

5.3 Specific Humanities Ontologies

This series comprises more specific entities, as used in different specialized domains in the Humanities, e.g., about scholasticism in medieval philosophy. Some ontologies (e.g., *indology*, *catholic organization* and *philosophy*) are for the moment only providing the more general concepts as needed for the current projects, but they can, and will, be extended. *Catholic orders* and *philosophies* describe subclasses of classes in aforementioned ontologies, which will also be extended.

Although the scope of these ontologies is narrower, i.e., more project-oriented, the entities can be reused in another context, if applicable, meaning that they do not need to be restricted to a specific project.

External terminology and code systems ontology This ontology contains formal descriptions of terminology and code systems, and their datatypes, as links between such systems' data, OWL-ontologies and RDF-data. Examples of terminology and code systems in the Humanities are the ISO[20] standard OAIS[21], and the GND[22] for the DACH countries. A generic system, for example, is using hexadecimal color coding to describe e.g., text color. Other datatypes are declared in respective domain ontologies. Examples are `calendar:julianDate` in the calendar-ontology, to type a Julian date literal, `languages:iso639-2` in the languages-ontology, to type ISO language standard codes, and `text:characterSize` in the text-ontology, to type a numeral representing character size.

Project ontologies These ontologies contain entities that are only used in their respective projects, but which are still usable outside those projects, if applicable. For example, the Parzival-ontology contains the concept of *Dreissiger*, (being a set of about 30 verses,) which would be reusable in another project about the verse novel *Parzival*.

[20] International Organization for Standardization. Cf. https://www.iso.org.

[21] Open Archival Information System. Cf. http://www.oais.info/.

[22] Gemeinsame Normdatei. Cf. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html.

All other generic ontologies are, to different degrees, re-used in the project ontologies by creating subclasses or subproperties from definitions in the generic ontologies.

Generally, another project about a same author could reuse some ontological elements in order to be linkable with one of the projects supported by NIEINE. It then would be possible to query the two different SPARQL endpoints simultaneously, which is essential for research, since the triple stores can contain complementary information on the same subject. This is actually the case for the DRCS project (dealing with the commentaries on the *Sentences* of Peter Lombard), which is linked to another project at the University of Baltimore, U.S. This case demonstrates the added value of semantic interoperability between disparate databases, facilitated by the use of the same external ontologies CIDOC-CRM and FRBRoo.

Atharvaveda-ontology This ontology represents the formal description of specific concepts in the critical edition of the Paippalāda Recension of the Indian Atharvaveda (UZH 2020), an anonymous collection of verse songs about everyday life (c. 1200-1000 BC). It contains subclasses and subproperties of elements in the text, text expression, text editing, prosodic structure, literature, and Indology ontologies, thus, modelling a sub-set of the Indian Veda literature.

Delille-ontology This ontology represents the formal description of specific concepts in the scientific edition of the third canto of Jacques Delille's (1738–1813) verse poem *L'homme des champs* (*The Rural Philosopher*) (Marchal 2020). One of the main goals of the project is the reconstitution of the reception of Delille's poetry. The main concepts concern poetic expression and its structures, the works (e.g., anthologies, dictionaries, and articles) and their authors (and other actors) that cite verses of the canto, and the scientific commentaries on the citers and their works.

Dietrich-ontology This ontology represents the formal description of specific concepts in the critical edition about father Joseph Dietrich's monastery diary (1670-1704) (WSU 2020). The concepts are diary-related. The majority of the semantics are expressed using more generic ontologies.

DRCS-ontology This ontology represents the formal description of specific concepts in the scientific study Digital Repertory of Commentaries on Peter Lombard's *Sentences* (DRCS, Peter Lombard c.1096–1160) (Zahnd 2020). The project collected over 1700 commentaries, from which further data has been extracted, such as the identification of the authors, their names and life dates, their membership of religious orders, their philosophical thinking and tradition, their roles in the editing processes (e.g., editor, abbreviator, corrector), along with all graspable information about the text itself, like genre, location and time of creation, as well as bibliographic information. The specific Humanities ontologies containing those concepts are shown in Figure 4. The DRCS-ontology itself mostly contains the different types of commentaries and the Stegmueller-related concepts. Interlinking this collected data through our ontologies provides highly dynamic possibilities for evaluating and detecting as-yet-unknown patterns, interrelations and networks in medieval philosophy and intellectual history. The reception of texts and writers, thought patterns, writing traditions or thought methods can be detected and traced down through time.

Kuno Raeber-ontology This ontology represents the formal description of specific concepts in the online publication of the lyric work of the Swiss poet Kuno Raeber (1922–1992) (Morgenthaler 2020). As the first model of a project-ontology in the learning curve of NIE, it contains more than the average number of multiparent subclasses. On the other hand, there is also the need for single instance classes leading to a more extensive ontology, than for other projects. The concepts mainly concern the different expression formats (written, typed, etc.), their carriers, and their convolutes, strongly representing the FRBRoo ontology.

Lavater-ontology This ontology represents the formal description of specific concepts in the critical edition of correspondence of Johann Caspar Lavater (1741–1801) (UZH Deutsches Seminar 2020). All concepts in this ontology are letter-related, as subclasses of more generic classes, enabling the internal structure to be kept close to other sources, as in the example e-manuscripta.ch.

Meyer-ontology This ontology represents the formal description of specific concepts in the critical edition of the correspondence of Conrad Ferdinand Meyer (1825–1898) (Lukas 2020). As in the Lavater-ontology, above, in this ontology all the concepts are letter-related, as subclasses of more generic classes; this occurs on the level of expressions and information carriers, as defined by FRBRoo.

Parzival-ontology This ontology represents the formal description of specific concepts in the critical and digital edition of the verse novel *Parzival* by Wolfram von Eschenbach (c. 1160-c. 1220) (Stolz 2020). One part of the concepts describes a series of transcriptions (or parts) of the verse novel, and corresponds therefore with the levels of expressions and information carriers in FRBRoo; the other part of the concepts describes the critical edition with different apparatus. This can be achieved by making extensive use of the scholarly editing ontology.

Wölfflin-ontology This ontology represents the formal description of specific concepts in the scientific study Heinrich Wölfflins gesammelte Werke (1864–1945) (UZH Kunsthistorisches Institut 2020). Due to the current absence of a consolidated source data model, the ontology is not yet published. However, it should be noted that most of the required semantics is already covered in more generic ontologies.

The Kritische Robert Walser-Ausgabe (1878–1956) (UNIBAS Deutsches Seminar 2020) and the Anton Webern Gesamtausgabe (1883–1945) (UNIBAS Musikwissenschaftlichen Seminar 2020) govern their own ontologies, which they intend to link to our ontologies at a later stage.

5.4 Expected Results

We will further develop application-independent SWT, enabling semantic interoperability and reusability adhering to the FAIR-principles. The ontology library will grow, and the generic ontologies will stabilize, but still be extendable. We will increasingly model to allow for broader usage in the Humanities, beyond scholarly editing, and in consensus with aforementioned parties. The ontologies will be declared with enhanced expressiveness, considering the whole of RDF/S and OWL ontological elements. We will also start implementing machine-reasoning in different ways, to enrich research data with inferred domain knowledge; temporal reasoning will be an important part of this. We will have more ontological graphics, and extend the documentation, particularly regarding our modelling methodology, workflow, and best practices.

6 Conclusion

Being able to connect different research projects using formal semantics is more a consequence of the implementation of the SWT standards and modelling, than the primary intention of research projects. Of the eleven editions we support, only DRCS initially expressed the need for semantic interoperability (in this case with another US project with the same subject).

Even if researchers tend to focus on the specificities of their topics and objects, they have an important part of the basic semantics of research projects in common. It, therefore, makes sense to invest a little more effort to obtain broadly reusable models, which capture these common semantics. Such models, which come about as the result of consensus (without loss of expressivity), and which are based on standards of W3C and the domain of Humanities (e.g. CIDOC-CRM and derivatives), are necessary in order to avoid project-specific semantic silos, in which the same concepts are modelled over and over again for different projects, in a multiplication of effort and cost. Only in this way is the need for tedious reverse engineering minimized.

In our experience, researchers on an individual project quickly become aware of the advantages of *RDF-izing* their data, from the quality assurance on different levels (due to the explicitness of the formal data), to performing machine-reasoning on that data to answer research questions. Domain specialists don't have to dive into SWT themselves to understand its functionality and potential. We are aware that the following is a bold statement, but we are convinced that with RDF-data, one can infer new data in a way that is impossible with other data models, because of the lack of the built-in logic of the formal languages RDF/S, OWL and Notation3 (N3).

Although the notion of interoperability is included in the FAIR-principles, the understanding of semantic (machine) interoperability is growing slower. It is not enough to use the SWT standards and adopt *upper* ontologies. Consensus about domain semantics at different levels of specificity is indispensable. The willingness of domain specialists to discuss terminology in order to obtain a critical mass of consensus models is growing due to the aforementioned direct RDF advantages. Although domain specialists do have discussions among themselves to reach a semantic consensus, this consensus remains on the level of human understanding and is not explicitly formalized in the data models to be machine-interpretable and reusable. This is why there is a real need for SWT experts who are able to liaise between domain specialists and IT staff, in order to bring the semantic consensus to a new level. Also, the creation of a collaborative online environment to create a library of ontologies has been very helpful, e.g., the NIE-INE GitHub repository (NIE-INE 2019b).

The project terminology discussions concerning the formalization of domain knowledge, facilitated by a SWT expert who operates across various projects and domains, complies with the tendencies of *multidisciplinaritization* and Linked Open Data (LOD) promoted by politics and policies. Of course, the intellectual property rights aspects have to be dealt with appropriately.

Because the formal data are independent from natural language, foreign collaboration is also facilitated, internationalizing domain semantics.

Different projects that deal with the same topics are likely to cover different aspects that are complementary, and very worthwhile to connect on the formal semantic level, increasing the chance of new research findings. Here, the next step in the technological implementation comes into play: machine-reasoning based on N3-rules consuming RDF-data and OWL-ontologies, for semantic conversion and for answering research questions.

Bibliography

Allemang, Dean, and James Hendler, *Semantic Web for the Working Ontologist* (Oxford: Elsevier LTD, 2011)

Berners-Lee, Tim, 'The Semantic Web as a Language of Logic', 1998 <https://www.w3.org/DesignIssues/Logic.html>

Brickley, Dan, and Libby Miller, 'FOAF Vocabulary Specification 0.99', *FOAF Vocabulary Specification 0.99*, 2014 <http://xmlns.com/foaf/spec/>

CIDOC, 'CIDOC Documentation Standards Working Group, and CIDOC CRM SIG', 2006 <http://www.cidoc-crm.org/>

Ciotti, Fabio, 'A Formal Ontology for the Text Encoding Initiative', *Umanistica Digitale*, 3 (2018) <https://umanisticadigitale.unibo.it/article/view/8174>

Ciotti, Fabio, and Francesca Tomasi, 'Formal Ontologies, Linked Data, and TEI Semantics', *Journal of the Text Encoding Initiative*, 9 (2016) <https://doi.org/10.4000/jtei.1480>

DCMI, *Dublin Core Metadata Initiative Schemas*, 1995 <https://web.archive.org/web/20190930150058/https://www.dublincore.org/schemas/>

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear, 'What Is Text, Really?', *Journal of Computing in Higher Education*, 1.2 (1990), 3–26

Dunsire, Gordon, 'Functional Requirements for Bibliographic Records Object-Oriented Extension to the the CIDOC Conceptual Reference Model', 2014 <http://metadataregistry.org/schema/show/id/94.html>

Eide, Øyvind, and Christian-Emil Ore, 'Ontologies and Data Modeling', in *The Shape of Data in Digital Humanities* (Routledge, 2018)

ESIP, 'Official Repository for Semantic Web for Earth and Environmental Terminology (SWEET) Ontologies', 2019 <https://github.com/ESIPFed/sweet>

Gabler, Hans Walter, 'Theorizing the Digital Scholarly Edition', *Literature Compass*, 7.2 (2010), 43–56 <https://doi.org/10.1111/j.1741-4113.2009.00675.x>

IFLA Study Group, *Functional Requirements for Bibliographic Records*, IFLA Series on Bibliographic Control 19 (Munich, 1998) <https://web.archive.org/save/https://www.ifla.org/publications/functional-requirements-for-bibliographic-records>

Library of Congress, *Extended Date/Time Format (EDTF) Specification* (Library of Congress, 2019) <https://www.loc.gov/standards/datetime/>

Lukas, Wolfgang, 'C. F. Meyers Briefwechsel-Kritische Ausgabe', *C. F. Meyers Briefwechsel-Kritische Ausgabe*, 2020 <http://www.cfmeyer.ch/>

Marchal, Hugues, 'Reconstruire Delille', *Reconstruire Delille*, 2020 <https://delille.philhist.unibas.ch/>

Miles, Alistair, and Sean Bechhofer, 'SKOS Simple Knowledge Organization System Namespace Document', 2009 <https://www.w3.org/2009/08/skos-reference/skos.html>

Morgenthaler, Walter, 'Kuno Raeber Lyrik', *Kuno Raeber Lyrik*, 2020 <https://www.kunoraeber.ch/lyrik/>

NIE-INE, 'Nationalen Infrastruktur Für Editionen - Infrastructure Nationale Pour Les Éditions', *Nationalen Infrastruktur Für Editionen*, 2019a <http://e-editiones.ch/about>

———, 'NIE-INE Ontologies', 2019b <https://github.com/nie-ine/Ontologies>

Pierazzo, Elena, *Digital Scholarly Editing: Theories, Models and Methods* (Farnham, Surrey: Ashgate, 2015)

Renear, Allan H., David G. Durand, and Elli Mylonas, 'Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies', in *Research in Humanities Computing*, ed. by S. Hockey and N. Ide (presented at the ALLC/ACH Conference, Christ Church, Oxford, April 1992, Oxford: Oxford University Press, 1996) <http://cds.library.brown.edu/resources/stg/monographs/ohco.html>

Robinson, Peter, 'Towards a Theory of Digital Editions', *The Journal of the European Society for Textual Scholarship*, Variants, 10 (2013) <https://doi.org/10.1163/9789401209021_009>

Stolz, Michael, ‘Parzival-Projekt’, *Parzival-Projekt*, 2020 <http://www.parzival.unibe.ch/home.html>

UNIBAS Deutsches Seminar, ‘Kritische Robert Walser-Ausgabe’, *Kritische Robert Walser-Ausgabe*, 2020 <https://kritische-walser-ausgabe.ch/>

UNIBAS Musikwissenschaftlichen Seminar, ‘Anton Webern Gesamtausgabe’, *Anton Webern Gesamtausgabe*, 2020 <https://anton-webern.ch/index.php?id=17>

UZH, ‘Online Edition of the Paippalāda Recension of the Atharvaveda’, *Online Edition of the Paippalāda Recension of the Atharvaveda*, 2020 <https://www.atharvavedapaippalada.uzh.ch/en.html>

UZH Deutsches Seminar, ‘Lavater-Edition’, *Edition Johann Caspar Lavater*, 2020 <https://www.lavater.uzh.ch/de.html>

UZH Kunsthistorisches Institut, ‘Heinrich Wölfflin - Gesammelte Werke’, *Heinrich Wölfflin - Gesammelte Werke*, 2020 <https://www.woelfflin.uzh.ch/de.html>

Verborgh, Ruben, and Jos De Roo, ‘Drawing Conclusions from Linked Data on the Web: The EYE Reasoner’, *IEEE Software*, 32.3 (2015), 23–27 <https://doi.org/10.1109/MS.2015.63>

W3C, ‘Notation3 (N3): A Readable RDF Syntax’, *Notation3 (N3): A Readable RDF Syntax*, 2011 <https://www.w3.org/TeamSubmission/n3/>

———, ‘OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)’, *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)*, 2012 <https://www.w3.org/TR/owl2-syntax/>

———, ‘OWL Web Ontology Language Overview (W3C Recommendation 10 February 2004)’, *OWL Web Ontology Language*, 2004b <https://www.w3.org/TR/2004/REC-owl-features-20040210/>

———, ‘Product Modelling Using Semantic Web Technologies’, *Product Modelling Using Semantic Web Technologies*, 2009 <https://www.w3.org/2005/Incubator/w3pm/XGR-w3pm-20091008/>

———, ‘RDF 1.1 Primer’, *RDF 1.1 Primer*, 2014a <https://www.w3.org/TR/rdf11-primer/>

———, ‘RDF 1.1 Turtle RDF Triple Language’, *RDF 1.1 Turtle*, 2014c <https://www.w3.org/TR/turtle/>

———, ‘RDF Primer (W3C Recommendation 10 February 2004)’, *RDF Primer*, 2004a <https://www.w3.org/TR/rdf-primer/>

———, ‘RDF Schema 1.1’, *RDF Schema 1.1*, 2014b <https://www.w3.org/TR/rdf-schema/>

———, ‘SPARQL 1.1 Overview’, *SPARQL 1.1 Overview*, 2013 <https://www.w3.org/TR/sparql11-overview/>

———, ‘W3C Semantic Web’, *Semantic Web*, 2001 <https://www.w3.org/standards/semanticweb/>

Wilkinson, Mark D., Michel Dumontier, and I. Aalbersberg, ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship’, *Scientific Data*, 3 (2016) <https://doi.org/10.1038/sdata.2016.18>

WSU, ‘Das Kloster-Tagebuch Des Einsiedler Paters Joseph Dietrich, 1670–1704’, *Das Kloster-Tagebuch Des Einsiedler Paters Joseph Dietrich, 1670–1704*, 2020 <http://www.dietrich-edition.unibe.ch/>

Zahnd, Ueli, ‘A Digital Repertory of Commentaries on Peter Lombard’s Sentences’, *A Digital*

Repertory of Commentaries on Peter Lombard's Sentences, 2020 <https://drcs.zahnd.be/index.php>

The Critical Apparatus Ontology (CAO): Modelling the TEI Critical Apparatus as a Knowledge Graph

Francesca Giovannetti

Abstract

This paper seeks to explore the use of semantic web technologies to enhance the re-/presentation of the critical apparatus that accompanies a TEI digital scholarly edition. The apparatus is a key instrument for critical editions. Its encoding poses a challenge for researchers, who strive to achieve highly expressive digital representations of their scholarly views. So far, no comprehensive ontology has been developed for the representation of the critical apparatus. This study makes a first step towards filling this gap by proposing the Critical Apparatus Ontology, an OWL ontology for representing the critical apparatus as a knowledge graph.

1 Introduction

This study proposes a conceptual model for the representation of the TEI critical apparatus as a knowledge graph. Critical apparatuses are fundamental tools for scholarly editions dealing with works that exist in multiple versions; they provide a window on the editor's workshop, and offer readers the evidence they need in order to evaluate the edition itself; evidence includes, but is not restricted to, variant readings.

When modelling or using a critical apparatus, it is of key importance to understand its purpose: a critical apparatus does not provide a mere list of textual variants, but rather presents textual variants in a way capable of conveying the editor's theory about how the readings and witnesses are related to one another (Damon 2016). In other words, a critical apparatus should put textual variants into context, conceptually sitting at the focal point of the network of texts that participate in the process of reconstruction of a work. Every element of a critical apparatus should be approached as the result of an act of scholarly interpretation (Romanello et al. 2009).

Representing the critical apparatus in TEI can be challenging. It is an operation which involves structuring the critical text and apparatus as hierarchies of ordered, nesting, XML elements. However, as Sperberg-McQueen points out, textual variation represents "a type of textual non-linearity" which hardly fits into a tree data structure (Sperberg-McQueen 1989). This paper argues that the use of a graph model – and of knowledge graphs specifically – instead of a tree model, provides a more intuitive structure for representing a critical apparatus.

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 125–139.

The Resource Description Framework (RDF) is a graph data model for capturing and representing knowledge as statements composed of a *subject*, a *predicate*, and an *object* which, when linked to one another and supported by formal semantics (i.e. by one or more ontologies), form a knowledge graph. A number of studies have begun to discuss the use of ontologies to provide TEI-encoded documents with a more formal semantics (e.g., Ciotti & Tomasi 2016; Eide 2014/15; Jordanous et al. 2012; Ciula et al. 2008). Several scholars have recommended that more digital scholarly editions should be published according to the principles of linked open data (LOD), and in collaboration with cultural heritage institutions that are already making their collections available in the LOD cloud (e.g. Daquino & Tomasi 2015; Ore & Eide 2009). Nonetheless, most digital scholarly editions remain document-centric, thus missing out on the opportunities and the greater gains that come with a data-centric approach.

The benefits of using knowledge graphs, in support of TEI XML encoding, to enhance digital scholarly editions outweigh the production costs involved:

1. Entities can be linked to one another via meaningful links to form networks. TEI provides two methods for describing a relationship between elements: one is using specific XML attributes (e.g., `<rdg wit="A">` to express that the reading is witnessed in A); the other is nesting (e.g., `<app><rdg>a</rdg><rdg>b</rdg></app>` to indicate that reading "a" and "b" are alternative readings). However simple, there always remains a margin of ambiguity when interpreting such relationships, as semantics has to be deduced from structure. Knowledge graphs provide a machine-readable method for representing complex connections and their semantics, including textual variation, situations where fragments of the same witnesses are scattered across different libraries, and text transposition (which involves a relationship across distinct locations of the text and, as such, is particularly problematic to represent in a document-centric standard).
2. Entities can be represented according to different, interlinked, perspectives. For example, the concept of reading could be described as the result of a scholarly interpretation, or as a fragment of text.
3. The provenance of an entity or graph can be easily specified by means of dedicated ontological vocabularies. This facilitates the process of assessing information quality.
4. An open-ended number of different, and even conflicting, arguments can be expressed (in such a case, declaring the provenance of information is central).
5. Knowledge graphs offer concrete opportunities for federation. Siloed editions are dismantled as soon as references to external, centralized, repositories are provided. For example, a critical apparatus may relate the edition to an external linked data record of one of the witnesses (e.g., a record held by a cultural institution, such as a library or museum). In this way, interoperability between editions, and the

inclusion of digital scholarly editions in the cultural heritage-linked open data cloud, becomes possible.

6. Ontologies provide the content model with a good degree of flexibility, as users are able to introduce new classes and property without compromising interoperability; the structure of a TEI document can vary on a case-by-case basis, while RDF, with its graph structure, remains consistent across applications.

Some of these objectives are achievable even in an entirely TEI-based environment. However, the use of knowledge graphs allows a more straightforward formalization of complex, interrelated theories and interpretations.

The Critical Apparatus Ontology (CAO) is a conceptual model providing a framework for the representation of the critical apparatus as a knowledge graph. CAO may be used in combination with TEI documents to enhance the edition with formal semantics. An experimental Python script is available for converting specific features of an existing TEI critical apparatus to a CAO knowledge graph.[1]

2 Related Works

Previous attempts to represent the phenomenon of textual variation in the form of graphs have led to the development of automatic collation tools such as CollateX (Dekker et al. 2015), whose underlining data structure is the variant graph, i.e. a graph-oriented model for the representation of textual variants, originally developed by Desmond Schmidt (Schmidt & Colomb 2009). Other projects, more concerned with visualization, have moved in a similar direction (e.g. Andrews & Macé 2013; Jänicke et al. 2015).

To date, there has been little work on standard ontologies for the description of the critical apparatus that accompanies an edition of a text. The Linking Ancient World Data (LAWD) ontology was developed for the purpose of enhancing interoperability between data about the ancient world gathered from different projects (Cayless 2015). The LAWD ontology supports the description of single variant readings. As such, expanding the LAWD model with new properties to represent specific relationships between variant readings could have been a viable option. However, LAWD lacks a way of distinctly representing the concept of reading as a scholarly interpretation, and the textual fragment supporting such a reading: a LAWD reading can be both a scholarly interpretation and a text, at the same time. The need for distinguishing the text from its interpretation motivates the choice of creating a new conceptual model for the description of the textual variants and the philological arguments which together form a critical apparatus.

[1] The transformation script can be downloaded from https://github.com/fgiovannetti/cao/tree/master/tei-to-cao.

Consistent with best practice, which encourages reuse of existing ontologies, the Critical Apparatus Ontology (CAO) is based on the TEI abstract model, and incorporates classes and relationships from several conceptual models to solve specific issues: the Web Annotation Data Model (Sanderson et al. 2017) is used for describing scholarly annotations and to link them to the critical text; specific properties of the PROV-O ontology (Lebo et al. 2013) are used to declare the provenance of the interpretations, and to describe specific types of relationship between textual variants; FRBRoo (Bekiari et al. 2015) is used to describe witnesses according to different layers of analysis (i.e., work, expression, manifestation and item).

3 An Overview of the Critical Apparatus Ontology

The Critical Apparatus Ontology is an OWL ontology for the representation of the critical apparatus that accompanies a digital scholarly edition. In line with the philosophy of the TEI (TEI Consortium 2019, Ch. 12), CAO does not imply commitment to any particular school of textual criticism.

This section illustrates how to represent the critical apparatus according to the model provided by CAO, from the creation of a new apparatus entry, to the description of the readings and their relationships, and to the formalization of the process of scholarly interpretation, which lays at the foundation of the development of any type of critical edition. The following section does not provide a full description of the model (for which, see the online specification),[2] but, rather, aims to illustrate its expressive potential through a number of examples.

4 Creating a New Scholarly Annotation

The creation of a new CAO scholarly annotation (which corresponds to the introduction of a new entry in the critical apparatus) is done using the Web Annotation standard.[3] Each scholarly annotation is composed of a *body* (the content of the apparatus entry) and a *target* (the locus of variation in the text). The Web Annotation standard allows the declaration of the date of creation and of the creator, making it possible to attach provenance information to every single scholarly annotation.

```
@base <http://example.org/> .
@prefix cao: <http://w3id.org/cao/> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dcterms: <http://w3id.org/dc/terms/> .
@prefix frbroo: <http://iflastandards.info/ns/fr/frbr/frbroo/> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
```

[2] The full specification of CAO is available at http://w3id.org/cao.

[3] The same approach is adopted by LAWD (see the section “Related works”).

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .[4]

<annotation/anno01> a oa:Annotation ;
  oa:hasBody <app/app01> ;
  oa:hasTarget <varloc/vl01> ;
  dcterms:creator <person/fgiovan> ;
  dcterms:created "2019-03-09"^^xsd:date .
```

5 The Body of the Annotation

5.1 The Apparatus Entry

Suppose, for example, that the scholarly annotation we just created above aims to present the reader with a set of three possible variant readings for a specific locus of variation within the critical text.

The class `cao:VariationUnit` represents a cluster of variant readings or unit of variation. The property `cao:hasReading` links a unit of variation (`cao:VariationUnit`) to *n* variant readings. Therefore, in this example, the RDF description of the unit of variation would be as follows:

```
<app/app01> a cao:VariationUnit ;
  cao:hasReading <rdg/anno01-rdg01> ,
    <rdg/anno01-rdg02> ,
    <rdg/anno01-rdg03> .
```

5.2 The Readings

An apparatus entry may feature two distinct types of variant readings: generic readings, equivalent to the TEI element `<rdg>` (represented by the class `cao:Reading`), and base readings, equivalent to the TEI element `<lem>` (`cao:BaseReading`). The class `cao:BaseReading` is a subclass of `cao:Reading`; it therefore inherits its characteristics. A unit of variation may relate to a maximum of 1 base reading. The text of each reading is introduced by the property `rdf:value`. Continuing our example:

```
<rdg/anno01-rdg01> a cao:BaseReading ;
  rdf:value "diffundi"^^xsd:string.

<rdg/anno01-rdg02> a cao:Reading ;
  rdf:value "diffudit"^^xsd:string .

<rdg/anno01-rdg03> a cao:Reading ;
  rdf:value "diffundit"^^xsd:string .
```

[4] Prefixes are declared only once at the beginning of this section, but they apply to all examples.

5.3 Classifying the Readings by Type and Cause

Textual variants can be classified by category. In TEI, the `@type` attribute used on the elements `<rdg>` or `<lem>` serves this precise function. For example, an editor may classify a variant as an omission, a conjecture, an addition, or as a copyist's mistake. CAO represents reading types as individuals of the class `cao:ReadingType` (which is a subclass of `crm:E55_Type`). Possible values include:

- addition (the type for a reading that is an addition, i.e., a syntactic and/or semantic expansion);
- conjecture (the type for a reading that is a conjecture, i.e., an editorial reconstruction or hypothesis of reading);
- correction (the type for a reading that is a correction, i.e., an authorial or scribal correction of a previous textual fragment);
- deletion (the type for a reading that is a deletion, i.e., text marked or somehow indicated by the author or scribe as deleted);
- omission (the type for a reading that is an omission (i.e. the author or scribe did not include the reading in the witness; if intentional, an omission may also classify as a correction);
- transposition (the type for a reading that was transposed from another location in the text, including the flipping of the order of two words or phrases).

Wherever possible, each reading type has been aligned with the SKOS classification scheme for readings provided by LAWD.

For reading causes, encoded in TEI using the `@cause` attribute, CAO provides a new classification scheme which includes:

- dittography (the unintentional repetition of a letter, word or phrase);
- haplography (the accidental omission of a letter, word, or phrase; homeoarchy, homeoteleuton, and saut du même au même are particular types of haplography);
- homeoarchy (the accidental skipping of a word or phrase having the same beginning; a polyptoton is a particular type of homeoarchy);
- polyptoton (the accidental skipping of a word or phrase forming a polyptoton, i.e., presenting the same root word);
- homeoteleuton (the accidental skipping of a word or phrase having the same ending);
- saut du même au même (the accidental skipping of some text in between two similar words or phrases);
- incorporation (the accidental incorporation of materials, such as marginalia).

The properties `cao:hasReadingType` and `cao:hasReadingCause` are used to relate a `cao:Reading` to a type, and a cause, respectively. For example, the RDF description of an omitted reading would be as follows:

```
<rdg/anno01-rdg03> a cao:Reading ;
  cao:hasReadingType cao:omission ;
  cao:hasReadingCause cao:homeoteleuton ;
  rdf:value ""^^xsd:string .
```

5.4 Relating the Readings to One Another

The possibility of defining semantic relationships among entities is one of the greatest advantages of graph data modelling. CAO features various types of relationship between readings and, on a case-by-case basis, it is easy to extend the model to accommodate all kinds of unforeseen connections. For example, a chain of authorial revisions may be described as shown below using the property `cao:correctedTo` to identify the relationship between a reading and its modified version, forming a sequence of corrections. So, the following TEI-encoded apparatus entry

```
<app xml:id="anno08">
  <rdg varSeq="1" xml:id="anno08-rdg01">
    <del>this</del>
  </rdg>
  <rdg varSeq="2" xml:id="anno08-rdg02">
    <del><add>such a</add></del>
  </rdg>
  <rdg varSeq="3" xml:id="anno08-rdg03">
    <add>a</add>
  </rdg>
</app>
```

would become:

```
<app/app08> a cao:VariationUnit ;
  cao:hasReading <rdg/anno08-rdg01> ,
    <rdg/anno08-rdg02> ,
    <rdg/anno08-rdg03> .

<rdg/anno08-rdg01> a cao:Reading ;
  cao:correctedTo <rdg/anno08-rdg02> ;
  cao:hasReadingType cao:deletion .

<rdg/anno08-rdg02> a cao:Reading ;
  cao:correctedTo <rdg/anno08-rdg03> ;
  cao:hasReadingType cao:deletion .

<rdg/anno08-rdg03> a cao:Reading ;
  cao:hasReadingType cao:addition .
```

Other CAO properties to describe the relationships among readings are: `cao:follows`, `cao:hasVariant`, `prov:wasDerivedFrom`, and `prov:wasRevisionOf`.

In TEI, the `@varSeq` attribute conveys information about the chronological sequence in which variants have appeared over time. In the same way, the property `cao:follows` is used to chain the readings to link the readings together in chronological sequence.

The property cao:hasVariant links a cao:BaseReading to its alternative readings (which are members of the superclass cao:Reading). The property prov:wasDerivedFrom denotes "a transformation of an entity [i.e. a reading] into another, an update of an entity resulting in a new one, or the construction of a new entity based on a pre-existing entity" (Lebo et al. 2013).

The property prov:wasRevisionOf is used for specific cases of derivation of a reading from another one: it is a subproperty of prov:wasDerivedFrom and defines "a derivation for which the resulting entity is a revised version of some original. The implication here is that the resulting entity contains substantial content from the original" (Lebo et al. 2013). The domain and range are instances of the class cao:Reading. Below, are some examples:

```
<rdg/anno01-rdg02> a cao:Reading ;
  cao:follows <rdg/anno01-rdg03> .

<rdg/anno01-rdg02> a cao:Reading ;
  prov:wasDerivedFrom <rdg/anno01-rdg03> .

<rdg/anno01-rdg02> a cao:Reading ;
  prov:wasRevisionOf <rdg/anno01-rdg03> .
```

5.5 Citing Other Scholars' Readings

In TEI, the @source attribute is used on <rdg> to indicate responsibility for the claim that a witness supports a particular reading. The @source attribute contains a reference to an external source, normally a published edition. In a fully developed LOD scenario, it would be possible to navigate from one critical edition to another via the links between readings, which would act like bridges across editions. In CAO, a cited reading is related to its original source via the property prov:hadPrimarySource. For example:

```
<rdg/anno01-rdg02> a cao:Reading ;
  rdf:value "diffudit"^^xsd:string ;
  prov:hadPrimarySource <http://example.org> .
```

5.6 From the Reading to the Witness

Texts reach us in multiple versions. Each textual version represents a different realization of the same work. As anticipated in the introduction, CAO reuses concepts from the FRBRoo ontology to represent a text according to different levels of analysis. An FRBR expression, i.e., a text, is modelled using the class frbroo:F2_Expression, which defines "the specific intellectual or artistic form that a work takes each time it is realized" (IFLA 2009). A reading is a scholarly claim: the outcome of an act of interpretation, which is supported by a particular fragment of a witness, a `frbroo:F23_Expression_ Fragment`.

The `cao:witnessedBy` property relates a reading to an expression, i.e., to the text of the witness (either the full text, or the specific textual fragment that supports the reading within a witness), as shown in the example below:

```
<rdg/anno01-rdg01> a cao:BaseReading ;
  rdf:value "diffundi"^^xsd:string ;
  cao:isWitnessedBy <wit-fragment/anno01-wfrag01> ,
  <wit-fragment/anno01-wfrag02> .

<rdg/anno01-rdg02> a cao:Reading ;
  rdf:value "diffudit"^^xsd:string ;
  cao:isWitnessedBy <wit-fragment/anno01-wfrag03> .

<rdg/anno01-rdg03> a cao:Reading ;
  rdf:value "diffundit"^^xsd:string ;
  cao:isWitnessedBy <wit-fragment/anno01-wfrag04> .
```

The property `frbroo:R15i_is_fragment_of` connects an expression fragment to the expression to which it belongs:

```
<wit-fragment/anno01-wfrag01> a frbroo:F23_Expression_Fragment ;
  frbroo:R15i_is_fragment_of <wit-expression/wexp01> .

<wit-fragment/anno01-wfrag02> a frbroo:F23_Expression_Fragment ;
  frbroo:R15i_is_fragment_of <wit-expression/wexp02> .
```

Expressions cannot exist without a carrier. For example, a handwritten note could not exist without the piece of paper on which it is written. A carrier is, in FRBR terms, a manifestation. FRBRoo distinguishes between two classes of manifestation: Manifestation Singleton and Manifestation Product Type. A Manifestation Singleton belongs to the realm of physical things: it is a unique, physical object such as a manuscript, and cannot be replicated. On the other hand, a Manifestation Product Type is an abstract notion: it comprehends publications, which can exist in multiple physical copies.

Among the type of witnesses that a scholar may encounter are manuscripts, either handwritten or digital, which fall into the category of manifestation singletons, and printed editions, which belong to the class of manifestation product types. The properties `crm:P128i_is_carried_by` and `frbroo:R4_carriers_provided_by` relates an Expression to a Manifestation Singleton and a Manifestation Product Type, respectively:

```
<wit-expression/wexp01> a frbroo:F2_Expression ;
  crm:P128i_is_carried_by <wit/wit01> .

<wit-expression/wexp01> a frbroo:F2_Expression ;
  frbroo:R4_carriers_provided_by <wit/wit02> .
```

At this stage, possibilities of connection with other datasets open up. There are, indeed, two main ways to describe the witnesses involved in the reconstruction of a critical text: describing each witness within the model in detail, or relying on external descriptions, such as those provided by museums, cultural institutions, and other datasets (e.g., the British Museum, Pleiades, GeoNames, Worldcat, VIAF, DBpedia). CRM-FRBRoo is fully equipped for the description of manuscripts as museum artefacts, as well as for publications.

For example, the URI for the object of the property `crm:P128i_is_carried_by` may link to information belonging to external cultural heritage datasets: the URI itself can be directly reused from already existing records; alternatively, project-specific URIs can be paired with external representations of the same objects by means of the property `owl:sameAs`, as follows:

```
<wit/wit01> a frbroo:F4_Manifestation_Singleton ;
  owl:sameAs <http://collection.britishmuseum.org/id/object/exampleID> .
```

To sum up, a reading (`cao:Reading`) is a scholarly interpretation of a witness fragment (`frbroo:F23_Expression_Fragment`) that belongs to a witness (`frbroo:F2_Expression`), which is carried by a physical document (`frbroo:F4_Manifestation_Singleton`) or an edition (`frbroo:F3_Manifestation_Product_Type`). Other types of witnesses may also be described by using other elements of FRBRoo.

6 A Special Case of Variation: Transposition

Transposition occurs when text is transferred from one location to another. As transposition involves a relationship between a transposed text and (at least) two distinct locations, its representation requires that the annotation be linked to multiple targets:

```
<annotation/anno02> a oa:Annotation ;
  oa:hasBody <app/app02> ;
  oa:hasTarget <varloc/vl02>, <varloc/vl03> ;
  dcterms:creator <person/fgiovan> ;
  dcterms:created "2019-03-09"^^xsd:date .
```

The transposed text is then related to the original and the new location via the properties `cao:wasTransposedFrom` and `cao:wasTransposedTo`, respectively:

```
<rdg/anno02-rdg02> a cao: Reading ;
  cao:hasReadingType cao:transposition ;
  cao:wasTransposedFrom <varloc/vl02> ;
  cao:wasTransposedTo <varloc/vl03> .
```

7 The Target of the Annotation: Linking the Apparatus to the TEI Document

The example provided below shows a method for linking the apparatus to a specific base text within the TEI document containing the edition. The Web Annotation class `oa:RangeSelector` allows the identification of the beginning and of the end of the location of the variant using XPath expressions:

```
<l>
... freta rapidisque <anchor xml:id="varloc-s01"/>diffundi<anchor xml:id="varloc-
     e01"/> ...
</l>

<varloc/vl01> a cao:VariationLocation ;
  oa:hasSource <example.xml> ;
  oa:hasSelector [
    a oa:RangeSelector ;
      oa:hasStartSelector [
        a oa:XPathSelector ;
        rdf:value "//l/anchor[@xml:id='varloc-s01']" ] ;
        oa:hasEndSelector [
          a oa:XPathSelector ;
          rdf:value "//l/anchor[@xml:id='varloc-e01']" ] ] ] .
```

XPointer may alternatively be used to specify the location of the variant:

```
<varloc/vl01> a cao:VariationLocation ;
  oa:hasSource <example.xml> ;
   oa:hasSelector [
     a oa:FragmentSelector ;
      dcterms:conformsTo <http://tools.ietf.org/rfc/rfc3023> ;
      rdf:value "xpointer(//anchor[@xml:id='varloc-s01']/range-to(//anchor
[@xml:id='varloc-e01'])" ] .
```

The TEI double-end-point-attached method solves the issue of overlapping lemmata. There are, however, other methods which can be used for linking an RDF apparatus to the text. For example, if the location of the variant is tagged using a generic TEI element such as `<seg>`, the apparatus entry can directly point to the unique identifier specified for that element:

```
<l>... freta rapidisque <seg xml:id="varloc-01">diffundi</seg> ...</l>
<varloc/vl01> a cao:VariationLocation ;
  oa:hasSource <example.xml> ;
   oa:hasSelector [
     a oa:XPathSelector ;
     rdf:value "//l/seg[@xml:id="varloc-01]" ] .
```

It is also feasible to replace the lemma with an empty element in the TEI document, leaving a gap in the critical text (this would allow the editor not to choose a base or preferred text). Such a method, however, may result in loss of information if the TEI document containing the edition and the RDF critical apparatus are not processed together.

A Web Annotation class `oa:TextPositionSelector` is also available for describing ranges of characters, making it possible to link an RDF critical apparatus to a plain text document. Using this method, however, would make the connection between the apparatus and the text more fragile as compared to using `@xml:ids` for building the URIs representing the location of variation.

For a more detailed overview of the Critical Apparatus Ontology, please visit the ontology specification at http://w3id.org/cao.

8 Generating, Visualizing, and Querying an RDF Critical Apparatus

This study set out to explore the use of a conceptual model, the Critical Apparatus Ontology (CAO), in combination with TEI text encoding to describe the critical apparatus, and discussed the benefits of this approach. However, there are practical and methodological issues such as how to generate, visualize, and query an RDF critical apparatus, which deserve consideration: the lack of user-friendly tools for working with RDF and ontologies is a barrier for the adoption and diffusion of Semantic Web technologies in the context of digital scholarly editing (Pierazzo 2016).

There are different ways to generate an RDF critical apparatus. A first method is to directly extract the set of RDF statements from an existing TEI critical apparatus by means of an existing Python script which employs lxml, a library for manipulating XML documents (see Behnel 2019), and RDFLib, a package for working with RDF (see RDFLib Team 2013). A limitation of this approach is that the TEI encoding model applied to the critical apparatus must follow a specific structure and set of guidelines; otherwise it will not be possible for the script to convert the critical apparatus in its entirety. For example, in case of author's or copyist's additions, the script requires
`<rdg xml:id="rdg032" type="cao:addition">some added text</rdg>`[5]
rather than

```
<rdg xml:id="rdg032">
  <add>some added text</add>
</rdg>
```

Another possible strategy for generating an RDF critical apparatus is represented by RDFa, which allows the embedding of RDF into TEI documents through attributes, and comes with ready-for-use extraction tools (see, for example, Tittel et al. 2018 and Ruiz et al. 2020).

Nonetheless, both options present the same limitation: there is not a standard way of encoding certain features of a critical apparatus, such as specific types of relationships between variant readings, with TEI.[6] Therefore, it is not possible to

[5] Note that any vocabulary for the classification of reading types is allowed, as long as its prefix is defined in a `<prefixDef>` element.

[6] The only type of relationships that have a standard method of TEI encoding (the `@varSeq` attribute) are generic sequential relationships, translated into CAO as `cao:follows` relationships.

automatically extract such features unless alternative encoding methods, such as the `<graph>` or `<relation>` elements, are used.[7]

That being said, the exercise of human judgement (and manual intervention) on the generated triples remains central to the development of reliable critical apparatuses.

Studies on visualization are, without a doubt, indispensable for enabling regular web users to make sense of the RDF data presented, and for the diffusion of the LOD paradigm. There are several existing tools for the visualization of semantic knowledge bases which could transform RDF critical apparatuses into meaningful graph visualizations (for example, see Bastian et al. 2009). Alternatively, ad hoc visualizations may be created by means of the aforementioned RDFLib (SPARQL queries are used to extract relevant information from the knowledge graph) or the more familiar XSLT. For example, we could imagine a base text or, in its absence, a user-selected text with points of variation highlighted, and made interactive. The user can ask for different kinds of visualizations, such as variant graphs, variants in context, correction processes, outgoing links, raw RDF statements, and so forth.

A web scholarly edition enhanced by means of semantic graphs should also provide a SPARQL endpoint to allow advanced searching and extraction of the data. End users should be able to compose queries in abstract terms, without needing to know any specific formal query language (Ciotti 2018).

9 Conclusion and Future Work

Representing the critical apparatus within a document-driven environment such as TEI XML can be difficult, as a critical apparatus sits at the intersection of multiple texts, and establishes a complex network of relationships among these texts. RDF is a data-driven modelling framework. Compared to XML, RDF is closer to the way natural language communication is structured. The adoption of RDF knowledge graphs in combination with TEI may: facilitate the development of software for the visualization of variant graphs; provide scholars with more powerful ways of expressing their analysis and interpretations; allow users to perform semantic queries; increase interoperability between digital scholarly editions and the cultural heritage linked open data cloud.

The Critical Apparatus Ontology (CAO), currently in its 0.9^{th} version, aims to contribute to the publication of data-centric digital scholarly editions. Future work will focus on two main areas: carrying out the necessary tests on existing editorial projects before a stable release of the model; and, research on future updates – such

[7] Encoding RDF triples within `<graph>` or `<relation>` represents a case of tag abuse. However, if TEI truly is "whatever you make it" (Cummings 2012), a widespread tag misuse might quickly become a standard.

as the representation of reading types as classes rather than individuals – which will increase the expressiveness of the model.

Bibliography

Andrews, Tara L., and Caroline Macé, 'Beyond the Tree of Texts: Building an Empirical Model of Scribal Variation through Graph Analysis of Texts and Stemmata', *Literary and Linguistic Computing*, 28.4 (2013), 504–521

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy, 'Gephi: An Open Source Software for Exploring and Manipulating Networks', in *Third International AAAI Conference on Weblogs and Social Media*, 2009

Behnel, Stefan, *Lxml - Processing XML and HTML with Python*, 2019 <https://lxml.de/4.3/>

Bekiari, Chryssoula, Martin Doerr, Patrick Le Bœuf, and Pat Riva, *Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism* (IFLA, 2015)

Cayless, Hugh, *Linked Ancient World Data (LAWD) Ontology* (Linked Ancient World Data Initiative (LAWDI), 2015) <https://github.com/lawdi/LAWD>

Ciotti, Fabio, 'A Formal Ontology for the Text Encoding Initiative', *Umanistica Digitale*, 2.3 (2018)

Ciotti, Fabio, and Francesca Tomasi, 'Formal Ontologies, Linked Data, and TEI Semantics', *Journal of the Text Encoding Initiative*, 9 (2016) <https://doi.org/10.4000/jtei.1480>

Ciula, Arianna, Paul Spence, and José Miguel Vieira, 'Expressing Complex Associations in Medieval Historical Documents: The Henry III Fine Rolls Project', *LLC*, 23 (2008), 311–25 <https://doi.org/10.1093/llc/fqn018>

Cummings, James, 'TEI, What Else?', 2012 <https://prezi.com/_y7t4nulanba/tei-what-else/>

Dadzie, Aba-Sah, and Matthew Rowe, 'Approaches to Visualising Linked Data: A Survey', *Semantic Web* 2,2 (2011), 89–124 <http://semantic-web-journal.net/content/approaches-visualising-linked-data-survey>

Damon, Cynthia, 'Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts', in *Digital Scholarly Editing*, ed. by Matthew James Driscoll and Elena Pierazzo, Theories and Practices, 1st edn (Open Book Publishers, 2016), ɪᴠ, 201–18 <https://www.jstor.org/stable/j.ctt1fzhh6v.15>

Daquino, Marilena, and Francesca Tomasi, 'Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects', in *Metadata and Semantics Research*, ed. by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou, Communications in Computer and Information Science (Springer International Publishing, 2015), 424–36

Dekker, Ronald Haentjens, Dirk Van Hulle, Gregor Middell, Vincent Neyt, and Joris van Zundert, 'Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project', *DSH*, 30 (2015), 452–70 <https://doi.org/10.1093/llc/fqu007>

Eide, Øyvind, 'Ontologies, Data Modeling, and TEI', *Journal of the Text Encoding Initiative*, Issue 8 (2014/15) <https://doi.org/10.4000/jtei.1191>

Giovannetti, Francesca, *The Critical Apparatus Ontology (CAO)*, 2019 <http://w3id.org/cao>

IFLA, *Functional Requirements for Bibliographic Records: Final Report* (IFLA, 2009) <https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>

Jänicke, Stefan, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony, and Gerik Scheuermann, 'TRAViz: A Visualization for Variant Graphs', *DSH*, 30 (2015) <https://doi.org/10.1093/llc/fqv049>

Jordanous, Anna, K. Faith Lawrence, Mark Hedges, and Charlotte Tupman, 'Exploring Manuscripts: Sharing Ancient Wisdoms across the Semantic Web', in *WIMS*, 2012 <https://doi.org/10.1145/2254129.2254184>

Lebo, Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, and others, *Prov-o: The Prov Ontology*, 2013 <http://www.w3.org/TR/prov-o/>

Light, Michelle, and Tom Hyry, 'Colophons and Annotations: New Directions for the Finding Aid', *The American Archivist*, 65.2 (2002), 216–30 <https://doi.org/10.17723/aarc.65.2.l3h27j5x8716586q>

Ore, Christian-Emil, and Øyvind Eide, 'TEI and Cultural Heritage Ontologies: Exchange of Information?', *LLC*, 24 (2009), 161–72 <https://doi.org/10.1093/llc/fqp010>

Pierazzo, Elena, *Digital Scholarly Editing: Theories, Models and Methods - Elena Pierazzo - Libro in Lingua Inglese - Taylor & Francis Ltd - Digital Research in the Arts and Humanities| IBS* (Routledge, 2016) <https://www.ibs.it/digital-scholarly-editing-theories-models-libro-inglese-elena-pierazzo/e/9781472412119>

RDFLib Team, *RDFLib Documentation*, 2013 <https://rdflib.readthedocs.io/en/stable/>

Romanello, Matteo, Monica Berti, Federico Boschetti, Alison Babeu, and Gregory R. Crane, 'Rethinking Critical Editions of Fragmentary Texts by Ontologies', in *ELPUB*, 2009

Ruiz Fabo, Pablo, Helena Bermúdez Sabel, Clara Isabel Martínez Cantón, Elena González-Blanco García, and Borja Navarro Colorado, 'The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings', in *DSH*, 2020 <https://doi.org/10.1093/llc/fqaa035>

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young, *Web Annotation Data Model. W3C Recommendation*, 2017

Schmidt, Desmond, and Robert Colomb, 'A Data Structure for Representing Multi-Version Texts Online', *International Journal of Human-Computer Studies*, 67.6 (2009), 497–514 <https://doi.org/10.1016/j.ijhcs.2009.02.001>

Sperberg-McQueen, C. Michael, 'A Directed-Graph Data Structure for Text Manipulation', 1989 <http://cmsmcq.com/1989/rhine-delta-abstract.html>

TEI Consortium, *P5: Guidelines for Electronic Text Encoding and Interchange* release 3.6.0, 2019 <https://tei-c.org/guidelines/>

Tittel, Sabine, Helena Bermúdez-Sabel, and Christian Chiarcos, 'Using RDFa to Link Text and Dictionary Data for Medieval French', in *W23 - 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science*, May 12, 2018 Phoenix Seagaia Conference Center Miyazaki, Japan, ed. by John P. McCrae, Christian Chiarcos, Thierry Declerck, Jorge Gracia, Bettina Klimek (presented at the LDL-2018, Myazaki, Japan, 2018), 7–12 <http://lrec-conf.org/workshops/lrec2018/W23/pdf/10_W23.pdf>

Projects and Editions

Transforming TEI Manuscript Descriptions into RDF Graphs

Toby Burrows, Matthew Holford, David Lewis, Andrew Morrison, Kevin Page, Athanasios Velios

Abstract

This paper reports on the transformation of the Bodleian Library's online medieval manuscripts catalogue from XML documents into RDF graphs. The catalogue uses the "Manuscript Description" section of the TEI Guidelines to encode entries which were originally published in printed form, but also incorporates amendments and additions from unpublished documents. The transformation of this catalogue has required the development of processes to extract the relevant elements from the TEI XML documents, assemble these extracts into a new XML file, and match the various elements and attributes to CIDOC-CRM and FRBRoo entities and properties which can be expressed as RDF triples and incorporated into graph databases. As a result of this work, information from the manuscripts catalogue has been ingested into Linked Data graph databases developed by two Oxford projects: OXLOD (Oxford Linked Open Data) and MMM (Mapping Manuscript Migrations).

1 Introduction

This paper reports on the transformation of the Bodleian Library's online medieval manuscripts catalogue, based on the Text Encoding Initiative (TEI) Guidelines, into RDF graphs using the CIDOC-CRM and FRBRoo ontologies, which enable integration of datasets. The catalogue uses the Manuscript Description section of the TEI Guidelines to encode entries which were originally published in printed form, but also incorporates amendments and additions from unpublished documents prepared in the Bodleian Library. The transformation of this catalogue has required the development of processes to extract the relevant elements from the TEI XML documents, assemble these extracts into a new XML file, and match the various elements and attributes to CIDOC-CRM and FRBRoo entities and properties which can be expressed as RDF triples and incorporated into graph databases. A particular focus of this work has been the provenance data relating to these manuscripts.

As a result of this work, information from the manuscripts catalogue has been ingested into Linked Data graph databases developed by two Oxford projects: (a) *OXLOD*, a pilot project on Linked Data in Oxford which brought together data from

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 143–154.

a range of Oxford's cultural institutions, and (b) *Mapping Manuscript Migrations* (MMM), an international project designed to track the ownership and provenance of medieval manuscripts using data from three databases related to manuscript history (Burrows et al. 2018), including Bibale, the Schoenberg Database of Manuscripts, and the Bodleian's manuscripts catalogue.

The technical infrastructure of these two projects is quite different: *OXLOD* used the ResearchSpace software developed by the British Museum and Metaphacts (Oldman & Tanase 2018), while MMM uses the Sampo-UI interface developed by the Semantic Computing Research Group at Aalto University (Hyvönen et al. 2019). But the projects share the same methodological approach based on Linked Open Data technologies like RDF triples, ontologies, and URIs. They both aim to integrate heterogeneous datasets relating to cultural heritage objects and to provide a more effective way of browsing and searching this kind of data.

2 State of the Art

The Manuscript Description section of the TEI Guidelines has become the de-facto schema for structuring detailed descriptions of manuscripts (TEI Consortium 2019).Institutions using TEI descriptions include the University of Pennsylvania (for its OPenn and Bibliophilly services) and Manuscriptorium, the digital manuscript library managed by the National Library of the Czech Republic. The Bodleian Library chose TEI for its online manuscript catalogues, launched in 2017 (Bodleian Library 2017). As well as the catalogue of medieval manuscripts in Oxford libraries, they include Fihrist, a cooperative catalogue of manuscripts from the Islamicate world, and seven other specialized Oxford catalogues.

The TEI Guidelines are designed to be flexible and hospitable to different approaches to encoding and markup. In the case of the Manuscript Description section, this means that there are various ways of encoding the same basic information, with no definitive agreed standards. The Bodleian Library has recently made available draft encoding guidelines for manuscript descriptions, aimed at promoting consistency of encoding across different catalogues (Bodleian Library 2019a). Developed in association with Cambridge University Library and the British Library, the guidelines for medieval manuscripts draw on earlier work by Patrick Granholm and Eva Nyström for the Swedish manuscript catalogue manuscripta.se.

In parallel, there has been growing interest in developing Linked Data approaches to the aggregation of manuscript data. Europeana, which includes records and images for more than 305,000 manuscripts, has established workflows for mapping manuscript records in a variety of formats to its Europeana Data Model (EDM), represented in RDF. The limitations of the EDM for manuscript metadata have been addressed by the

Digitised Manuscripts to Europeana (DM2E) project, which developed a specialization of the EDM to extend its properties and classes for use with manuscript records (Dröge et al. 2014). Some more specialized projects have also been applying CIDOC-CRM and FRBRoo to develop ontologies relating to manuscript descriptions, such as that summarized in Mancinelli et al. (2019).

Another important aggregator of manuscript data is Biblissima, which brings together descriptions from about forty mainly French manuscript catalogues. For its initial prototype, Biblissima combined data from the Mandragore and Initiale databases into an RDF-based framework, using an ontology modelled on CIDOC-CRM and FRBRoo (Frunzeanu et al. 2016). The full Biblissima service, however, is based on XML pivot tables instead, and includes mappings from the TEI. It uses Linked Data techniques to align data through external identifiers from services like data.bnf.fr, GeoNames, and VIAF, and also makes available an RDF version which can be queried through a SPARQL end-point (Robineau 2019).

The relationship between TEI manuscript descriptions and the world of Linked Data, RDF, and ontologies has only been given relatively limited attention. In 2007, Øyvind Eide and Christian-Emil Ore produced a draft mapping from TEI to CIDOC-CRM, covering a selection of "events, time appelations, actors and actor appelations" drawn from several areas of the TEI Guidelines, but not including the "Manuscript Description" section. The same two authors (Ore & Eide 2009) subsequently produced a set of recommendations for TEI extensions and adjustments aimed at making "the ontological information in a TEI document compliant with the other cultural heritage models" including CIDOC-CRM. They noted that Manuscript Description was one of the TEI sections where ontologically oriented elements are defined. Eide's more general reflections on linking the TEI with external ontologies can be found in Eide 2014. More recently, Ciotti and Tomasi (2016/17) and Ciotti (2018) have presented a model aimed at "furnishing the TEI with a semantics based on a formal ontology". Crompton and Schwartz (2018) have proposed the "the development of XSLT-backed tools to convert and connect otherwise incommensurable [TEI] data sets".

To our knowledge, the only previous work on transforming TEI-encoded manuscript descriptions into RDF has been carried out by the Medieval Electronic Scholarly Alliance (MESA), which is one of the nodes of the Advanced Research Consortium (ARC). The ARC RDF schema is designed for encoding descriptions of digital resources made available through the Collex interface, and consists of a number of Dublin Core elements supplemented by a few Collex-specific elements (Medieval Electronic Scholarly Alliance 2019a). Several TEI-based manuscript catalogues have been mapped for the MESA-Collex search interface, and one example of a transformation of a TEI manuscript description from the Walters Art Museum has been published (Medieval Electronic Scholarly Alliance 2019b). For the most part, the TEI elements involved are limited to title, language, and date, while the TEI

<provenance> element has simply been mapped to <dc:provenance> without any encoding or mapping of persons, places, or events within a provenance statement.

3 Methodologies

At the Bodleian Library a customized TEI schema is used. Written in the TEI's ODD schema language, it is available in RELAX NG, XSD and DTD versions, and is used in eight different manuscript catalogues managed by the Bodleian. Separate authority files for persons, organisations, places and works are maintained and linked to the medieval manuscript descriptions. The raw TEI-XML files are stored in a public GitHub repository, where they are grouped by manuscript collection (Bodleian Library 2019b). Publicly accessible and searchable versions of the files are made available through a Web site built with technologies including XSLT, xQuery, Solr and Blacklight.

TEI manuscript data can be complex, often describing manuscripts divided into several parts, each with its own history and containing works-within-works (e.g., a collection of poetry and individual poems). Information about the history and provenance of the manuscripts (the focus of the MMM project) has been encoded in different ways, such as a single XML element describing the entire history of the manuscript, or multiple <provenance> elements which each recount one event. Dates are encoded with date tags or attributes on the provenance element.

The first step in our workflow is to identify those parts of the TEI schema which will be needed to answer the research questions of the MMM project. An xQuery script is used to extract these parts and copy them into a simplified XML document. It also creates URIs for each included entity. This simplified XML output is then mapped to classes and properties of the CIDOC-CRM and FRBRoo ontologies using the 3M mapping tool (Oldman et al. 2010). CIDOC-CRM was chosen as the basis for the MMM data model because its event-oriented nature makes it well-suited to modelling the provenance events in manuscript histories. It was combined with FRBRoo in order to represent the works, expressions, and manifestations carried by manuscripts (Mapping Manuscript Migrations 2019). The mappings for the TEI <provenance> element are summarized in Table 1.

The Bodleian's XML authority files are handled as separate datasets following the same method. Manuscript instances are then integrated with the authority records via corresponding URIs. The records include references to URIs from external authorities such as the Virtual International Authority File (VIAF), GeoNames, the Getty Thesaurus of Geographical Names (TGN), Gemeinsame Normdatei (GND), and WikiData. These have been retained in the RDF output for the MMM project, where they have

TEI element	Field in simplified XML	Ontology mapping in 3M
Provenance	`provenance`	`crm:E5_Event`
	`provenance/@xml:id`	`URIorUUID`
	`provenance/text`	`crm:P3_has_note > Literal`
	`provenance/date`	`crm:P4_has_time-span > crm:E52_Time-Span`
	`provenance/org`	`crm:P11_had_participant > crm:E74_Group`
	`provenance/org[@role='formerOwner']`	`crm:P51_has_former_or_current_owner > crm:E74_ Group`
	`provenance/person`	`crm:P11_had_participant > crm:E21_Person and frbr:F10_Person`
	`provenance/person[@role='formerOwner']`	`crm:P51_has_former_or_current_owner > crm:E21_Person and frbr:F10_Person`
	`provenance/place`	`crm:P7_took_place_at > crm:E53_Place and frbr:F9_Place`

Table 1. Mappings for the TEI `<provenance>` element.

been used to match persons, places, and organizations with those present in the other two source datasets.

The time constraints of the OXLOD and MMM projects made it necessary to work with the existing TEI documents produced by the Bodleian Library. Within the project timeframes, it was simply not feasible to re-encode the files or to enhance the existing encoding manually. Future work might also include experimenting with parsing and extracting the unencoded narrative statements within a `<provenance>` element.

Nevertheless, some bulk updating was done to add generic provenance statements to multiple files, especially where the existing `<provenance>` elements for a specific named collection did not include an entry for the Bodleian's acquisition of the manuscripts from the named collector. These updates were done at the University of Pennsylvania Library by forking the relevant TEI documents from GitHub, writing a Ruby script to add standard `<provenance>` and `<change>` statements to each file, and returning the files to the Bodleian. The TEI documents for more than 2,000 Bodleian manuscripts have been enhanced in this way.

The ontology used to structure the transformation of the Bodleian data was subsequently used in developing the MMM unified data model. This model was derived by examining the data structures of the three contributing datasets. The result was a mixture of elements from CIDOC-CRM and FRBRoo, combined with a few entity classes and properties unique to MMM. While the MMM project took the Biblissima ontology into account, it was insufficient to handle all the MMM data, especially those relating to manuscript provenance and history, which are central to the aims of MMM but largely out of scope for Biblissima.

4 Discussion

While the TEI manuscript descriptions may appear to be highly structured, there are important elements within them which are not. The focus of the MMM project is on the history and provenance sections of the descriptions, which record the evidence for the production and ownership of manuscripts over the many centuries of their existence. To meet the requirements of this project, we needed to extract as much of this evidence as possible in a suitably structured form.

The TEI `<provenance>` elements, in particular, hold most of the information needed. But these elements are often presented as a free-text narrative with marked-up entities for those persons, organisations, dates and places mentioned in the narrative. They often also hold transcriptions of annotations and inscriptions on the manuscript itself.

This reflects the traditional approach to printed manuscript catalogues, where this kind of information is given primarily in narrative form. The TEI Guidelines, at least initially, were designed to encode digital versions of these printed catalogues. A typical example of the Bodleian Library's treatment of `<provenance>` encoding – for manuscript Lat. th. (Latin theology) d. 29 – looks like this:

```
<provenance>Titles (13-th15th cent.) 'Consuetudines Lanfranci et excerpta de
     poenitentiale. Item Seneca ad lucillum', and press-mark 'E 25' (fol. iii)</
     provenance>
<provenance><persName role="fmo" key="person_1259"><!-- not found. His tomb
     mentioned in VCH. --> William Brent of Willington, Warwick</persName>, given
      to <persName role="fmo" key="person_61575100">Sir William Dugdale</persName
     >, 'i Oct.1675' (fol.iii verso)</provenance>
<provenance>Thorpe catalogue, 1831, no. 4108</provenance>
<provenance>W. Shaw Mason (fol.iii verso)</provenance>
<provenance>4th Duke of Newcastle, Sotheby's 6 Dec. 1937, lot 955</provenance>
<provenance><persName role="fmo" key="person_40764209"> André de Coppet </
     persName>, his sale at Sotheby's 6 Dec. 1954, lot 34, bought by Quaritch for
      £95.</provenance>
<provenance>Given by <persName role="fmo" key="person_111104108"> Arthur. A.
     Houghton</persName>.</provenance>
```

Listing 1. TEI `<provenance>` encoding for Bodleian Library Lat. th. d. 29.

There are seven provenance statements here, most of which include an encoded personal name, usually that of a former owner. None of the dates have been encoded, however; nor have any of the booksellers' names.

As well as the `<provenance>` statements, we also made use of the `<origin>` encoding within the `<history>` element. This is more rigorously encoded, as the example for the same manuscript demonstrates:

```
<origin>
  <origDate calendar="Gregorian" notAfter="1200" notBefore="1150">
    12th century, second half
  </origDate>
  <origPlace>
    <country key="place_7002445">English</country>
  </origPlace>
</origin>
```

Listing 2. TEI `<origin>` encoding for Bodleian Library Lat. th. d. 29.

Here the date of production – usually given as an approximate verbal range – has been converted to Gregorian dates in a `notAfter` / `notBefore` pattern. The place of origin, whether a country or a more specific location, has been linked to the value in the Bodleian Library's authority file for places, which normally has an associated TGN identifier.

Our aim was to extract the salient information about history and provenance automatically from the narrative of transfers of ownership. By using a combination of role attributes relating to ownership (where available) and the encoded entities within the provenance statements, we were able to construct event-related statements linking the manuscript and the actors in its history. We limited our model to generic relationships of provenance activities (primarily ownership, acquisition, and production) to ensure the accuracy of the resulting RDF statements, rather than attempting to infer more specific relationships from narrative statements which lacked the necessary markup.

The rest of the required information for the MMM project is related to bibliographical descriptions of the manuscripts, which were also the focus of the OXLOD project. Titles of works, and their authors, have been consistently encoded in the TEI documents and linked to the relevant authority file entry, making them relatively straightforward to extract and match to FRBRoo entities and relationships. The other key piece of data from each TEI document is the manuscript `shelfmark` (Listing 3).

```
<msIdentifier>
  <country>United Kingdom</country>
  <region type="county">Oxfordshire</region>
  <settlement>Oxford</settlement>
  <institution>University of Oxford</institution>
  <repository>Bodleian Library</repository>
  <idno type="shelfmark">MS. Lat. th. d. 29</idno>
  <altIdentifier type="internal">
    <idno type="SCN">Not in SC (late accession)</idno>
  </altIdentifier>
</msIdentifier>
```

Listing 3. `<msIdentifier>` encoding for Bodleian Library Lat. th. d. 29.

The TEI treament of shelfmarks is highly structured and relatively straightforward to extract and transform into RDF triples. But a manuscript shelfmark is not the same as a unique identifier in the Linked Data sense of the term. To create a URI for each manuscript, the MMM project has reused the unique element of the Bodleian's URL for an individual manuscript, e.g., https://medieval.bodleian.ox.ac.uk/catalog/manuscript_1927 becomes http://ldf.fi/mmm/manifestation_singleton/bodley_manuscript_1927 in the MMM triple store. There is no global system of manuscript identifiers similar to ISBNs for books, though an International Standard Manuscript Identifier (ISMI) for Linked Data purposes has been proposed (Cassin 2018). The Bodleian Library is investigating the possible use of ARK identifiers for its manuscripts (Burns et al. 2019).

5 Results

The TEI schema, the xQuery script and the simplified XML output are all available from the Bodleian Library's GitHub repository, together with the 3M mapping file and the RDF representations. The MMM interface to query the RDF records became publicly available in January 2020: https://mappingmanuscriptmigrations.org/ The full MMM dataset can be downloaded from the Zenodo repository: https://doi.org/10.5281/zenodo.3667486 The OXLOD pilot project has not yet been made available for external access.

Our workflow output was initially evaluated through a series of SPARQL queries run against the OXLOD pilot data. These queries focused on identifying relevant manuscripts through origin, provenance and acquisition events, filtered by location and time period. Additional contextual information was supplied through federated queries on the Getty Thesaurus of Geographic Names (TGN) and on WikiData. Examples of these queries can be seen in Velios (2018).

A second evaluation was carried out for the MMM project. This involved running SPARQL queries against the aggregated RDF data from three source datasets (including the Bodleian catalogue), using a set of research questions identified by researchers connected with the project.

These queries have been able to produce results which match those obtained directly from the Bodleian site using a combination of keyword searches and browsing by persons and places. The RDF queries connected with places of production and ownership have been able to take advantage of the geographical hierarchies embedded in the Getty TGN, even though these are not explicitly present in the relevant Bodleian authority file. A query like "Find all manuscripts produced in Lombardy in the 15th century" will return manuscripts originating specifically from places like Milan, Brescia and Pavia, for instance. This is not possible using a single query in the native interface to the Bodleian catalogue, since each place has to be searched separately.

On the other hand, the RDF queries have identified some issues with interpreting the dates used in manuscript descriptions, which are often expressed in very approximate terms. A date range like "xv – xvi centuries" is encoded by the Bodleian by converting it to the Gregorian calendar:

```
<origDate calendar="Gregorian" notAfter="1599" notBefore="1400">
```

But should this manuscript be counted among those produced in the 15th century or not? The answer is a matter for the manuscript researcher rather than the TEI encoder, however, and should be defined in the SPARQL query independently of the TEI encoding.

The work done to transform the Bodleian documents for these two projects has demonstrated that it is possible to extract TEI-encoded manuscript data in a form which can be expressed as RDF, loaded to a graph database, incorporated into a Linked Data environment, and retrieved using SPARQL queries. But the nature of some of the TEI markup – and especially the lack of encoding for various components of the narrative `<provenance>` statements – means that the RDF representation cannot include all the relevant information from the catalogue records. In the Bodleian catalogue itself, the keyword search function can still find occurrences of (for example) a bookseller's name, even though this has not been encoded. Replicating this functionality in the RDF environment would mean either re-encoding the TEI files in a more thorough and structured way or developing additional scripts to parse, extract, and transform provenance information which is currently presented in unencoded narrative statements within a `<provenance>` element.

6 Future Work

An important output from the Mapping Manuscript Migrations project will be a set of recommendations for re-thinking the structure and encoding of the TEI `<provenance>` element to enable its more effective reuse in graph applications. These recommendations will draw on the concepts previously outlined by Ore and Eide (2009), but will also take into account the parallel work currently being done in the art museum and gallery community on documenting and reusing provenance information. This includes improving the structure of provenance records in museum databases (Bergen-Fulton et al. 2015), as well as transforming museum databases to Linked Data and RDF graphs based on CIDOC-CRM (Knoblock et al. 2017). The Linked Art Data Model, which is in the process of development, will have a specific section devoted to the provenance of art works, based on CIDOC-CRM as the core ontology (Linked Art Community 2019).

Our aim is to ensure that the project's recommendations relating to the TEI `<provenance>` element can be framed within the existing TEI Guidelines, and can be incorporated into the Bodleian Library's customization and encoding guidelines for medieval manuscripts. These improvements are not just a matter of improving the specifics of TEI encoding. They will also require a significant re-thinking of the way in which manuscript provenance information is recorded and structured within catalogue records.

Acknowledgements

The Bodleian Library's manuscripts catalogue redevelopment project was funded by the Mellon Foundation. The Mapping Manuscript Migrations project at the University of Oxford is funded by the UKRI Economic and Social Research Council under the Digging into Data Challenge of the Trans-Atlantic Platform. The OXLOD project was funded by Oxford University's IT Capital Plan as part of the GLAM Digital Strategy.

Bibliography

Berg-Fulton, Tracey, David Newbury, and Travis Snyder, 'Art Tracks: Visualizing the Stories and Lifespan of an Artwork' (presented at the MW2015: Museums and the Web 2015: the Annual Conference of Museums and the Web, April 8-11, 2015, Chicago, IL, USA, 2015) <https://mw2015.museumsandtheweb.com/paper/art-tracks-visualizing-the-stories-and-lifespan-of-an-artwork/>

Bodleian Library, 'Medieval Manuscripts in Oxford Libraries', 2017 <http://medieval.bodleian.ox.ac.uk/>

———, 'Official Repository for the Bodleian Libraries TEI-Based Western Medieval Manuscript Catalogue', 2019a <https://github.com/bodleian/medieval-mss>

———, 'TEI P5 Customization and Encoding Guidelines - Bodleian Library', 2019b <https://msdesc.github.io/consolidated-tei-schema/msdesc.html>

Burns, Halle, Toby Burrows, J. Stephen Downie, David Lewis, Kevin Page, and Athanasios Velios, 'Assessing the Practicality of ARK Identifier Usage in a Catalogue of Medieval Manuscripts' (presented at the iConference 2019, 31 March - 3 April 2019, Washington, DC, 2019) <http://hdl.handle.net/2142/103380>

Burrows, Toby, Eero Hyvönen, Lynn Ransom, and Hanno Wijsman, 'Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts', *Manuscript Studies*, 3.1 (2018), 249–52 <https://repository.upenn.edu/mss_sims/vol3/iss1/13>

Cassin, Matthieu, 'ISMI: International Standard Manuscript Identifier: Project of Unique and Stable Identifiers for Manuscripts' (Hamburg, 2018) <https://www.manuscript-cultures.uni-hamburg.de/files/mss_cataloguing_2018/Cassin_pres.pdf>

Ciotti, Fabio, ‘A Formal Ontology for the Text Encoding Initiative’, *Umanistica Digitale*, 3 (2018) <https://umanisticadigitale.unibo.it/article/view/8174>

Ciotti, Fabio, and Francesca Tomasi, ‘Formal Ontologies, Linked Data, and TEI Semantics’, *Journal of the Text Encoding Initiative*, 9 (2016/17) <https://doi.org/10.4000/jtei.1480>

Crompton, Constance, and Michelle Schwartz, ‘More Than “Nice To Have”: TEI-To-Linked Data Conversion’ (presented at the DH2018, Mexico City, 2018) <https://dh2018.adho.org/more-than-nice-to-have-tei-to-linked-data-conversion/>

Dröge, Evelyn, Julia Iwanova, and Steffen Hennicke, ‘A Specialisation of the Europeana Data Model for the Representation of Manuscripts: The DM2E Model’, in *Assessing Libraries and Library Users and Use: Proceedings of the 13th International Conference Libraries in the Digital Age (LIDA), Zadar, 16-20 June 2014* (presented at the 13th International Conference Libraries in the Digital Age (LIDA), Zadar: University of Zadar, 2014), 41–50

Eide, Øyvind, ‘Ontologies, Data Modeling, and TEI’, *Journal of the Text Encoding Initiative*, Issue 8, 2014 <https://doi.org/10.4000/jtei.1191>

Eide, Øyvind, and Christian-Emil Ore, ‘Mapping of TEI to CIDOC-CRM, Version 0.1’, 2007 <http://www.edd.uio.no/artiklar/tekstkoding/tei_crm_mapping.html>

Frunzeanu, Eduard, Régis Robineau, and Elizabeth MacDonald, ‘Biblissima’s Choices of Tools and Methodology for Interoperability Purposes’, *CIAN-Revista de Historia de Las Universidades*, 19.1 <https://e-revistas.uc3m.es/index.php/CIAN/article/viewFile/3146/1783>

Hyvönen, Eero, Esko Ikkala, Jouni Touminen, Mikko Koho, Toby Burrows, Lynn Ransom, and others, ‘A Linked Open Data Service and Portal for Pre-Modern Manuscript Research’ (presented at the Digital Humanities in the Nordic Countries 2019 Conference, Copenhagen, 2019) <http://ceur-ws.org/Vol-2364/20_paper.pdf>

Knoblock, Craig A., Pedro A. Szekely, Eleanor E. Fink, Duane Degler, David Newbury, Robert Sanderson, and others, ‘Lessons Learned in Building Linked Data for the American Art Collaborative’, in *The Semantic Web – ISWC 2017*, Lecture Notes in Computer Science, Vol. 10588 (presented at the 16th International Semantic Web Conference, Vienna: Springer, 2017), Part II, 325–40

Linked Art Community, ‘Data Model: Provenance’, *Linked Art*, 2019 <https://linked.art/model/provenance/>

Mancinelli, Tizina, Antonio Montefusco, Sara Bischetti, Maria Conte, Agnese Macchiarelli, and Marcello Bolognari, ‘Modelling a Catalogue: Bilingual Texts in Tuscan Middle Ages (1260–1430)’, 2019 <https://dev.clariah.nl/files/dh2019/boa/1219.html>

Mapping Manuscript Migrations, ‘Data Model’, *Mapping Manuscript Migrations*, 2019 <https://github.com/mapping-manuscript-migrations/mapping-manuscript-migrations.github.io/tree/master/data_model>

Medieval Electronic Scholarly Alliance, ‘RDF Samples’, *Medieval Electronic Scholarly Alliance*, 2019a <http://wiki.collex.org/index.php/RDF_samples#MESA:_Walters_Art_Gallery>

———, ‘Submitting RDF’, *Medieval Electronic Scholarly Alliance*, 2019b <http://wiki.collex.org/index.php/Submitting_RDF>

Oldman, Dominic, and Diana Tanase, ‘Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace’, in *The Semantic Web – ISWC 2018. 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceed-*

ings, Part II, ed. by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, and others, Lecture Notes in Computer Science 11137 (presented at the ISWC 2018, Cham: Springer International Publishing, 2018), 325–340 <https://doi.org/10.1007/978-3-030-00668-6_20>

Oldman, Dominic, Maria Theodoridou, and Georgios Samaritakis, 'Using Mapping Memory Manager (3M) with CIDOC CRM. Version 4g', 2010 <http://83.212.168.219/DariahCrete/sites/default/files/mapping_manual_version_4g.pdf>

Ore, Christian-Emil, and Øyvind Eide, 'TEI and Cultural Heritage Ontologies: Exchange of Information?', *LLC*, 24 (2009), 161–72 <https://doi.org/10.1093/llc/fqp010>

Robineau, Régis, 'Biblissima: Connecting Manuscript Collections', 2019 <https://www.slideshare.net/biblissima/biblissima-connecting-manuscripts-collections>

TEI Consortium, *P5: Guidelines for Electronic Text Encoding and Interchange* (TEI Consortium, 2019) <https://tei-c.org/guidelines/>

Velios, Athanasios, 'Mapping MMM Data' (presented at the OXLOD Open Workshop 12 June 2018, Oxford, 2018) <https://www.glam.ox.ac.uk/sites/default/files/glam/documents/media/d10-7_ow7-slides_20180612_1.pdf>

Scholarly Music Editions as Graph: Semantic Modelling of the Anton Webern Gesamtausgabe

Stefan Münnich, Thomas Ahrend

Abstract

This paper presents a first draft of the ongoing research at the Anton Webern Gesamtausgabe (Basel, CH) to apply RDF-based semantic models for the purpose of a scholarly digital music edition. A brief overview of different historical positions to approach music from a graph-theoretical perspective is followed by a list of music-related and other RDF vocabularies that may support this goal, such as *MusicOWL*, *DoReMus*, *CIDOC CRMinf*, or the *NIE-INE* ontologies. Using the example of some of Webern's sketches for two drafted Goethe settings (M306 & M307), a preliminary graph-based model for philological knowledge and processes is envisioned, which incorporates existing ontologies from the context of cultural heritage and music. Finally, possible use-cases, and the consequences of such an approach to scholarly music editions, are discussed.

1 Introduction

Despite the seemingly irrefutable opinion that musicology is regarded as a "delayed discipline" (Gerhard 2000) due to the complexity of its subject-matter, scholarly music editions made the step into the digital age more than a decade ago.[1] Projects such as *Freischütz Digital* (2012*), Beethovens Werkstatt* (2014), or the *Digital Mozart Edition* (2006) have achieved fundamental milestones in the field of encoding musical information in XML-based formats (e.g., the format of the *Music Encoding Initiative* – MEI[2]) and have made significant contributions to both the theory of scholarly editing

[1] For a constitutive overview about the notion and concept of Digital (Scholarly) Edition see Sahle 2013, Pierazzo 2015, Driscoll and Pierazzo 2016, or Bleier et al. 2018. The diversity of the music philological landscape was cartographed in Emans and Krämer 2015; an in-depth discussion about the very nature of the *music philological question* was provided in Urbanek 2013. The transfer and application of digital methods to music editions has been fundamentally – in theory and practice – stimulated, discussed and developed during the last two decades by Joachim Veit and members of his research team (e.g. Veit 2006, Veit 2010, Veit 2015). A most comprehensive summary of the history, present and prospective of digital music philology can be found in Kepper 2011, a short repositioning in Acquavella-Rauch 2019, and in Kepper and Pugin 2017.

[2] See project website and latest specification of MEI (v4) in Music Encoding Initiative 2017 and Music Encoding Initiative 2018.

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 155–180.

and digital applications. However, the question arises (not only in the musical field) whether a *digital edition* should not be more than the mere conversion of a text into an XML format. The representation of semantic relationships and links between different areas of the editions, the overlap-free representation of non-hierarchical content, as well as the connection and interlinking with external data sets, are only some of the challenges that such an approach struggles to meet. Therefore, as early as 2009, Johannes Kepper considered whether "directed graphs are the more suitable data structure for encoding (music) texts than tree structures (and thus XML)" (Kepper 2009, 220). In the field of scholarly textual editions, some efforts were made in this direction (Kamzelak 2016, Wettlaufer 2018, or many other papers in the present volume). In the context of scholarly music editions, however, such an approach has, to date, hardly been tested. In order to pursue this desideratum, the Anton Webern Gesamtausgabe (University of Basel, CH) has the aim of researching and testing the scientific application of graph-based semantic models, in terms of RDF vocabularies,[3] for the purpose of a digital music edition. In this paper, we present a preliminary draft of this ongoing research: in section 2 we give a short overview of different historical positions to approach music from a graph-theoretical perspective, and a discussion of existing music-related and other, helpful, vocabularies; in section 3 we introduce a graph-based model for philological knowledge and processes, which is under active development within the project, in close cooperation with the Digital Humanities Lab Basel (DHLab) and the Swiss-wide National Infrastructure for Editions (NIE-INE). Incorporating existing ontologies from the context of cultural heritage and music, the possible interplay of these models is demonstrated using the example of some of Webern's sketches for two drafted Goethe settings from the 1930's (M306 & M307). The last section discusses possible consequences for the self-understanding of scholarly music editions if philological processes are considered as graphs.

2 Music as Graph

The idea to approach music from a graph-theoretical perspective has increasingly attracted attention since the mid-2000s in different areas:

On the one hand, graph-theoretical reflections have been applied to music in order to make its mathematical benefits available for music-analytical purposes.

[3] The *Resource Description Framework* (RDF) is a "standard model for data interchange on the Web. […] RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This *graph view* is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations." (RDF Working Group 2014).

Fundamental developments in the field of mathematical music theory were realized in geometric approaches (e.g., Mazzola 1990; Mazzola 2002; Tymoczko 2011) or especially in transformational approaches combining both group theory and graph theory (e.g., Lewin 1987; Lewin 1990; Klumpenhouwer 1998). Following on from these approaches, more current studies are covering a wide range of subjects, including pattern matching (e.g., Szeto & Wong 2006), musical gestures (e.g., Mazzola & Andreatta 2007), tonal modulation (e.g., Walton 2010), or voice-leading (e.g., Rings 2011; Popoff et al. 2018).

On the other hand, graph-based knowledge models of music have been developed which contribute to the vision of a semantic web as it was proposed by Tim Berners-Lee and others around the year 2000 (Berners-Lee 1998; Berners-Lee et al. 2001). Formalized by (RDF-based) ontologies in terms of "explicit, formal specification[s] of a shared conceptualisation" (Studer et al. 1998, 184), the aim of these models was to enable a machine-readable description and connection of music metadata, especially in the context of music information retrieval (MIR), music recommendation systems, or music library cataloguing.[4] Despite these overall efforts, only a few major international projects such as *DoReMus*[5] in France or *Transforming Musicology*[6] in the UK have promoted "the enhancement of Semantic Web provisions for musical study [...] augmenting existing controlled vocabularies (known as ontologies) for musical concepts"[7]. A comprehensive application of semantic web technologies to the modelling, enhancement, and transformation of human knowledge, as it has been discussed more and more in the humanities in recent years (e.g., Oldman et al. 2016), remains largely a desideratum in the domain of music,[8] and especially for scholarly music editions (Münnich 2018).

2.1 Existing Graph-Based Models for Musical Knowledge

When it comes to computer-based modelling of knowledge structures, it should be noted at the outset that each model can only be a reduced, simplistic, and imperfect *surrogate* for the considered part of the natural world, and that it can neither be all-

[4] A comparison of music metadata schemas is given in Corthaut et al. (2008); an overview of (graph-based) symbolic music representation systems can be found in Simonetta (2018).

[5] Project website of *DoReMus* (http://www.doremus.org/) and the data access point of the *DoReMus* project (http://data.doremus.org/).

[6] Project website of *Transforming Musicology* (https://tm.web.ox.ac.uk/). Nurmikko-Fuller and Page 2016.

[7] Description of the *Transforming Musicology* project on its earlier, now no longer accessible website: https://web.archive.org/web/20170225090608/http://www.transforming-musicology.org/about/.

[8] Daquino et al. 2017 surveyed the "Landscape of Musical Data on the Web" in 2017 and have published their findings as a Linked Open Dataset: https://github.com/enridaga/musow. Their observation is "that a large amount of [musical] resources are not ready to be part of the Web of Data", identifying "the heterogeneity of large collections, the uncertainty in licensing, and the lack of large scale approaches to semantic lifting of musical resources and data publishing" as the main obstacles (Daquino et al. 2017, 67). Many thanks to Albert Meroño-Peñuela for pointing us to this survey.

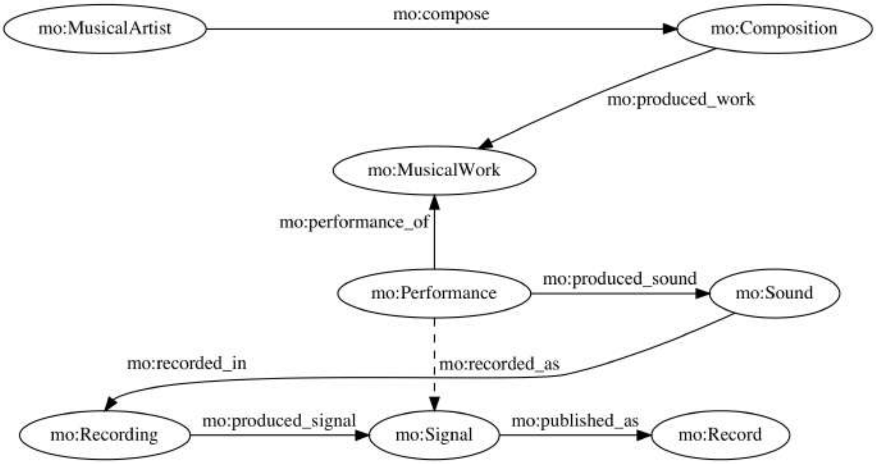


Figure 1. Excerpt (music production process) from the *MusicOntology*.

encompassing, nor finally conclusive (Davis et. al 1993; Stachowiak 1973). Thus, there is no single *correct* way, but rather multiple best possible ways of modelling, against which – up to a dead end, in case of doubt – the respective object of investigation, the questions to be applied to it, and one's own perspective, have to be tested.

Since as early as 2002, *MusicBrainz* has provided an online database of CD information that is stored and queried with unique identifiers (URIs) for artists, publishers, and albums, down to the track level. Although it does not offer any ontology in the narrower sense, *MusicBrainz* can be considered the first "semantic web service" for music-related information and it is still actively maintained and developed to date (Swartz 2002).

The *Music Ontology* (Raimond et al. 2013; motools 2013; motools 2007), which was developed at the Centre for Digital Music at the Queen Mary University of London in 2007, is widely used, especially in the field of MIR and music recommendation systems (Raimond et al. 2007; Sandler et al. 2009). Based on OWL, the Web Ontology Language, it is primarily concerned with statements on music production processes (like works, composers, performances, or recordings; see Figure 1). Its spin-offs and supplementary models (*timeline*, *event*, *keys*, *tonality*, *symbolic notation*, *chord*, *temperament*, *audio features*) allow the modelling and representation of further detailed musical information. Although it must currently be regarded as the de facto standard ontology for musical phenomena, the active development of the *Music Ontology* was discontinued in 2014.

The *MusicOWL – Music Score Ontology* (Jones et al. 2017a), which has been in devel-

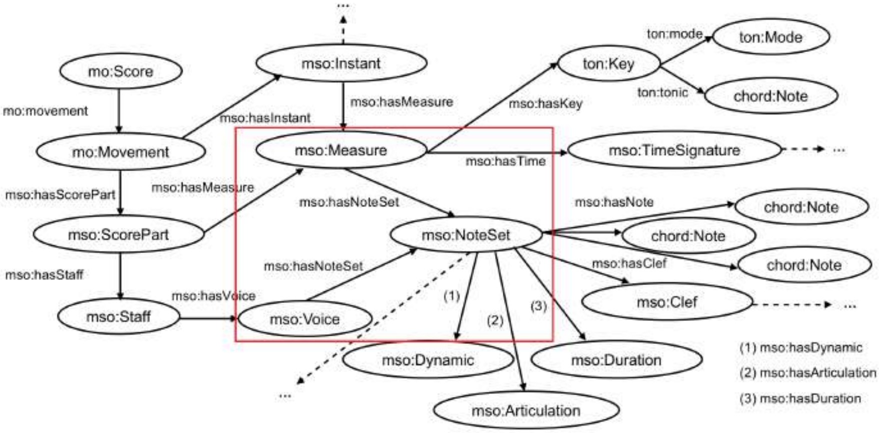


Figure 2. Excerpt from the *MusicOWL* ontology (mso) with integration of *Music Ontology* (mo), *Chord Ontology* (chord) and *Tonality Ontology* (ton).

opment at the University of Münster since 2017, is also based on OWL, but considers musical content beneath the level of musical scores, including elements such as parts, chords, notes, dynamics, articulations, measures, or voices (see Figure 2). Thereby, it reuses and expands the *Music Ontology* and its aforementioned supplementary models (Jones et al. 2017b). A very remarkable feature of the ontology is the overlap-free assignment of individual notes or chords, by means of a *NoteSet,* to both a certain measure or a certain part at the same time, which remains an intractable problem in XML-based encoding formats. Time will tell what the impact of *MusicOWL* will be, and that will be dependent on how well it handles or builds on existing standards. As an important first step, a (JAVA based) conversion tool from MusicXML to RDF, based on the *MusicOWL* ontology, already exists;[9] a corresponding module for the conversion of data encoded in MEI format would be most welcome in scholarly and philological contexts.

In addition, the *DoReMus* project has developed an ontology model that integrates and reuses the *MusicOntology*, *FRBRoo* (IFLA Working Group on FRBR/CRM Dialogue 2015), *CIDOC CRM* (CIDOC 2006), and the Europeana Data Model (EDM 2012), i.e., the main ontologies in the context of humanities (Achichi et al. 2015; Choffé & Leresche 2016). Thereby, a complex, mirrored, triangular modelling pattern, adapted from

[9] Github repository (https://github.com/jimjonesbr/musicowl). Another independent, but less documented approach to express MusicXML in RDF can be found in MusicML 2016. More information about the music interchange format MusicXML is provided in Good 2001. Latest specification in Good 2017.

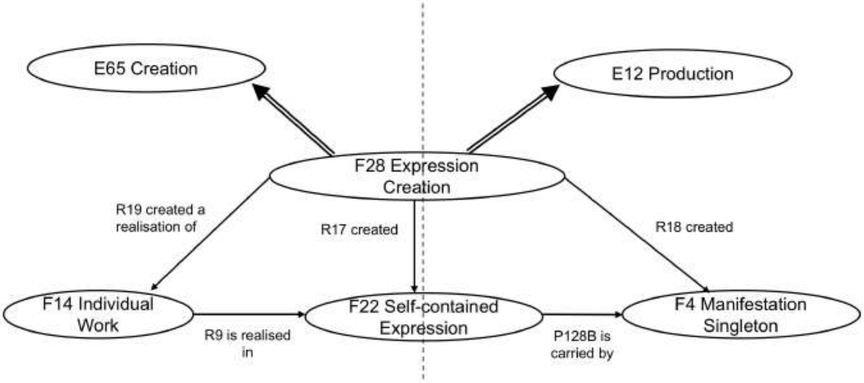


Figure 3. Mirrored triangular pattern in *FRBRoo.*

FRBRoo, comes into play (Figure 3): on a conceptual level, a self-contained expression (*F22*) of an individual work (*F14*) is created in an expression creation event (*F28*). On a physical level, a manifest sign carrier is created – the so-called manifestation singleton (*F4*) – that carries the content of the self-contained expression. This segmentation of the concept of a *work* enables a highly complex modelling pattern of, and differentiated statements about, the creation process of a composition, and enables the repetition of this pattern on the level of performances, publications, recordings, or reception processes (Figure 4).

With *MELD* ("Music Encoding and Linked Data")[10], the British *Transforming Musicology* project has created a semantic framework that researches the "distributed real-time annotation of digital music scores" (in MEI format) with the help of semantic technologies. Here too, several existing models are reused, including *Music Ontology*, *FRBR* (Functional Requirements for Bibliographic Records; IFLA Study Group 1998), *SKOS* (Simple Knowledge Organization System; Miles & Bechhofer 2009), *PROV-O* (Provenance Ontology; Lebo et al. 2013), and *Web Annotation Ontology* (Sanderson 2017). The goal is the enrichment of MEI data with Semantic Web Annotations, which should guarantee a dynamic real-time communication between the participants of a performance situation (orchestra or band members), mediated by the musical score (Weigl & Page 2017; Kallionpää et al. 2017). The *MELD* framework is utilized by the EU-funded *TROMPA* project ("Towards Richer Online Music Public-domain Archives")[11] that aims to establish a digital platform to interlink, digitize and semantically enrich

[10] Github-Repository *MELD* (https://github.com/oerc-music/meld).

[11] Project website *TROMPA* (https://trompamusic.eu/).

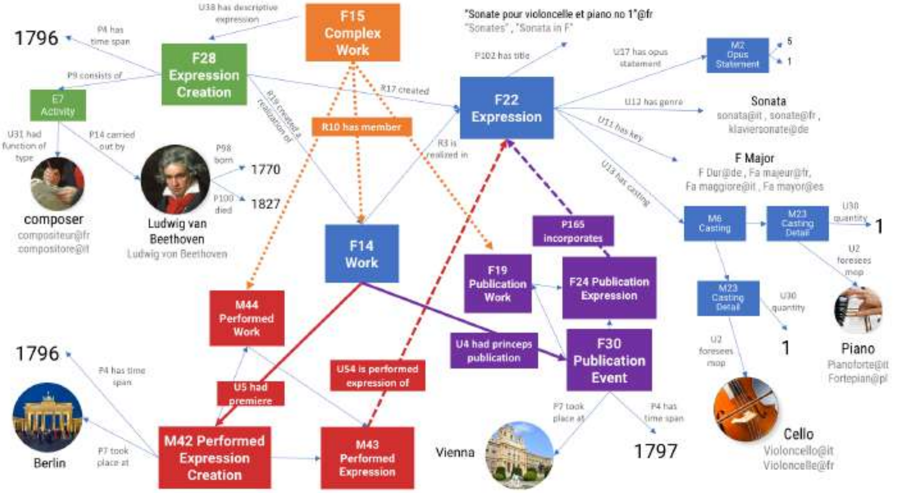


Figure 4. Excerpt from the *DoReMus* model with the separation of work (orange/blue), expression (blue) and expression creation (green) as well as their connection to performance (red) and publication (purple) processes by example of Beethoven's Cello sonata F-Major op. 5/1.

and annotate all forms of classical music in public domain, including scores in MEI format and sound recordings (Weigl et al. 2019; Goebl & Weigl 2019).

The *MIDI Linked Data* project researches the interconnection of "symbolic music descriptions [...] contained in MIDI files"[12]. Here, MIDI data are transformed to, and represented as, RDF graphs, with the help of a *MIDI Ontology* that captures the events (especially pitches, or instruments) of a MIDI file (Meroño-Peñuela & Hoekstra 2016a). A python-based converter allows for the lossless transformation from MIDI to RDF and back (Meroño-Peñuela & Hoekstra 2016b).

CHARM (Common Hierarchical Abstract Representation of Music) strives for an abstract representation of hierarchical musical structures. The concept developed in the early 1990s (Wiggins et al. 1989, Smaill et al. 1993) has been recently remodeled in a Semantic Web context (Harley & Wiggins 2015).

A no less ambitious project is *JazzCats* ("Jazz Collection of Aggregated Triples"),[13] which merges information on performances, recordings, and artists from three dif-

[12] Documentation of the *MIDI Linked Data* project, Github repository (https://github.com/midi-ld/documentation). Cf. also the project website: https://midi-ld.github.io/. The *Musical Instrument Digital Interface* (MIDI) is a control protocol for music devices and an industry standard since 1983. Its specifications can be found in MIDI Manufacturers Association (MMA) 2019.

[13] Project website *JazzCats* (http://jazzcats.cdhr.anu.edu.au/).

ferent jazz-related data sets (*Body & Soul*, *WJazzD*, and *LinkedJazz*), and links them together through concepts from the *Music Ontology* (Bangert et al. 2016; Bangert et al. 2018).

In 2016, the *Enhancing Music Notation Addressability* (EMA) project has introduced an Application Programming Interface (API) that facilitates "addressing and extracting specific portions of music notation published in machine-readable formats on the web" (Viglianti 2016, 57). Inspired by the URI-based mechanism and approach of the API of the *International Image Interoperability Framework* (IIIF) for images, the web service described by the Music Addressability API provides a standardized URI scheme, which can also be applied very effectively in graph modeling.

2.2 Other Helpful Models

In addition to generic top-level ontologies such as *FOAF* (Friend-of-a-Friend; Brickley & Miller 2014), *SKOS*, or the ontologies of the *Dublin Core Metadata Initiative* (DCMI 1995), which are extremely widespread due to the generality of their concepts, there are various models that are frequently used in connection with humanities, and that have been mentioned already in section 2.1: *CIDOC CRM* and the Europeana Data Model for Cultural Heritage (*EDM*), the bibliographic model *FRBRoo*, or *PROV-O* for provenance descriptions. Instead of these quite established models, we will concentrate in this section on two rather recent, and therefore less widely known, models: the first one is a whole series of ontologies that have been developed by the Swiss-wide project *National Infrastructure for Editions* (*NIE-INE*)[14]. These ontologies (Figure 5), adhering to the model theory of RDF and OWL Full, aim for a machine-interpretable, formal semantic expression of digital scholarly editions, providing a tremendous and highly interdependent range, from generic concepts (e.g., agent, event, human, organization), science-historical approaches (mathematics, philosophies, logic), to edition-specific (document, text-editing, text-structure, text, information carrier), or project-specific vocabularies (Kuno Raeber, Parzival, Atharvaveda, and others). Based on external ontologies (like the aforementioned) whenever possible, the *NIE-INE* ontologies apply event- and role-based modelling patterns (similar to, e.g., *CIDOC CRM*).

Another promising approach is the *CRMinf* Argumentation model, which was developed recently in the orbit of *CIDOC CRM* (Stead 2015; Doerr et al. 2015). It uses, inter alia, a pattern with a certain belief (*I2 Belief*), which assigns a corresponding truth value (*I6 Belief Value*) to a certain statement (*I4 Proposition Set*) as shown in Figure 6.

[14] Project website *NIE-INE* (https://www.nie-ine.ch) and Github-Repsitory *NIE-INE* (https://github.com/nie-ine). Authoritative publication of ontologies on http://e-editiones.ch. Cf. the paper of Roberta Padlina and Hans Cools in the present volume. The NIE-INE project was discontinued at the end of 2020.

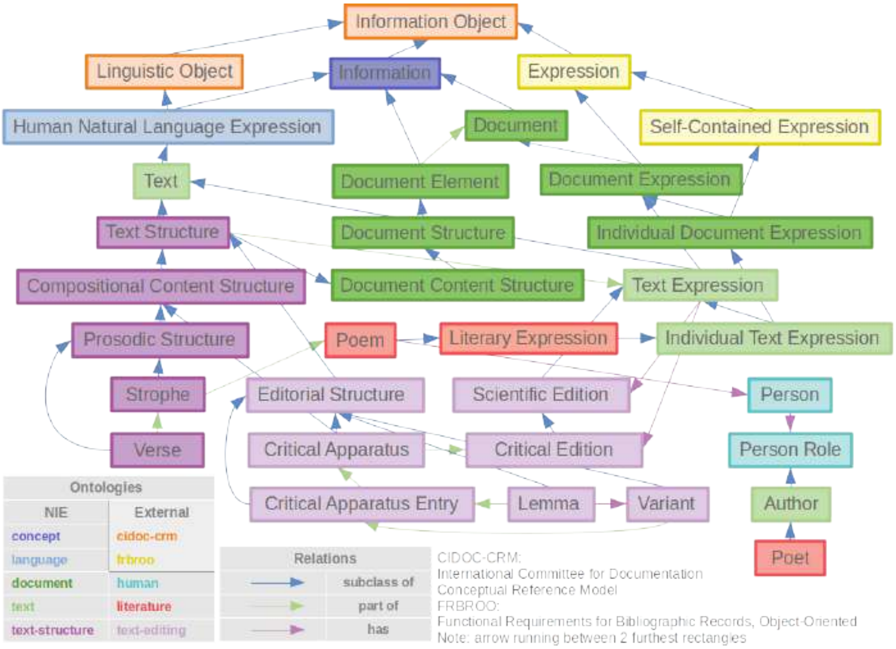


Figure 5. Core classes and properties of different ontologies from the NIE-INE project.

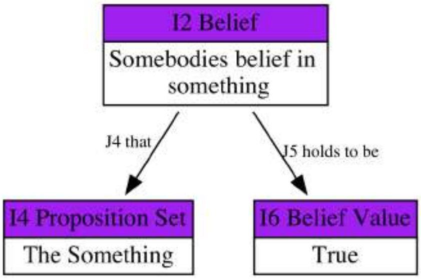


Figure 6. Belief Pattern in the *CRMinf* Argumentation Model.

This model could be particularly useful in the field of (digital) humanities, as it allows scholars to deal with uncertainty, doubts, hypotheses, or any kind of argumentative conclusion. However, so far only a few projects are known (like *ResearchSpace*[15], Oldman & Tanase 2018) that have applied this model comprehensively to an existing research question. Also, to gain the full potential of an argumentation model in the context of humanities, adding some time indicators to *CRMinf* should be considered ("I believe in something at a certain time"), as well as the possibility of using probability values, or at least a weighting of different beliefs.[16]

3 A Semantic Model for the Anton Webern Gesamtausgabe

The Anton Webern Gesamtausgabe (AWG 2015a; AWG 2015b) is working on a model that allows for a semantic representation of the philological knowledge compiled in the project (which is considered to be reusable for other music edition projects). As a so-called hybrid edition, located at the University of Basel, the AWG intends to make its digital parts available online with the help of the software framework *Knora* (Knowledge Organization, Representation, and Annotation), which is currently under development at the Digital Humanities Lab of the University of Basel.[17] (In addition, printed volumes will be published by Universal Edition in Vienna.) *Knora* allows one to supplement *Knora*-specific application models with project-specific ontology models of a certain granularity, which can be created, edited, and linked within the framework. In addition, facsimiles can be integrated, displayed, and annotated, in accordance with the IIIF standard (International Image Interoperability Framework Consortium 2019). For some time now, *Knora*'s predecessor, the virtual research environment *Salsah* (System for Annotation and Linkage of Sources in Arts and Humanities)[18], has been used productively as a database archive for context materials and for the document collection of the AWG. An area for daily editorial work, in which editors can create their critical reports and editions directly within the research environment, is currently under active development, also in close cooperation with the *NIE-INE* project (cf. section 2.2). A snapshot of the relationships between the various groups and projects involved is shown in Figure 7 in the form of an RDF graph, in which only classes or properties of the *FOAF*, *DCTerms* and *schema.org* (W3C Schema.org Community Group 2019) vocabularies are used. It goes without saying that the graph is not as limited as the figure conveys; according to the paradigm of an open world assumption, extensions and additions can be made at any point.

[15] Project website *ResearchSpace* (https://www.researchspace.org/index.html).

[16] Thanks to Hans Cools and Roberta Padlina for pointing this out.

[17] Project website *Knora* (https://www.knora.org).

[18] Project website *Salsah* (https://www.salsah.org/). Schweizer & Rosenthaler 2014.

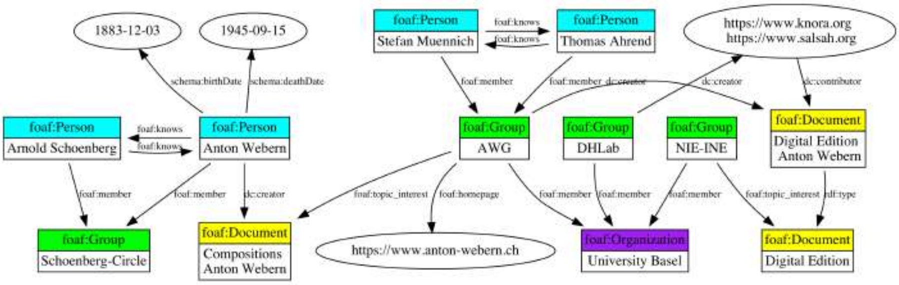


Figure 7. The Anton Webern Gesamtausgabe and some of its interrelations with historical and current groups, persons or matters of subjects.

Within this framework, the model of the AWG is being developed. Incorporating existing ontologies from the context of cultural heritage and music, this model will have a number of features: the separation of abstract works, self-contained expressions, and expression creation events (according to *FRBRoo*); the semantic embedding of these elements into their respective production, performance, or publication processes (according to *DoReMus*); the application of music-specific ontology models (*MusicOntology*, *MusicOWL*) and controlled vocabularies; as well as the integration of the *CIDOC CRMinf* Argumentation model, to deal with any kind of argumentative conclusion or uncertainty.

3.1 Graph-based model of philological knowledge and processes

Anton Webern's musical sketches for *Cirrus* M306, and for *Der Spiegel sagt mir: ich bin schön!* M307,[19] both unpublished fragments written in the summer of 1930 on poems by Goethe, shall serve as a starting point for the following discussion. Since M306 is conceptually a piano song, and M307 is, in most parts, a vocal composition, the two pieces are assigned to different sections of the edition (series II/3: Posthumous Choir Music, and series II/5: Posthumous Piano Songs). Each section contains its own general introduction, all the transcribed musical texts (sheets) of the section, and an overall critical report of the section (Figure 8). However, some sketches of M307 suggest a possible arrangement as a piano song, so the piece could be assigned to both sections. In most printed editions, assigning a piece to two different sections would hardly be conceivable, not least for reasons of space and cost. In a digital environment, however, the assignment does not have to be exclusive and can be designed to be flexible and multivariable.

[19] The M-number is referring to a cataloguing principle for Webern's oeuvre introduced by Hans Moldenhauer in Moldenhauer 1978.

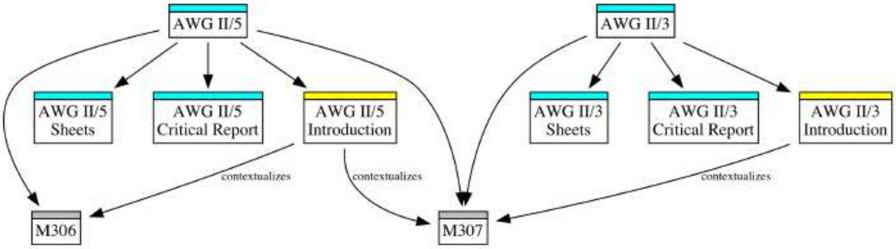


Figure 8. Assignment of different musical pieces (M306, M307) to different sections of the edition (AWG II/3 & II/5).

When looking at M307 a little closer (Figure 9), we see that the structure of the sections of the edition is mirrored on the level of the individual musical pieces: on a much more granular scale than the whole section, the musical pieces are equipped with a separate, more specific, introduction, their own specific transcribed musical texts (sheets), and a corresponding critical report. At the same time, these sub-sections are back-linked and contribute to their respective super-sections (indicated by dashed arrows). The critical report for M307 consists of an overview, a description, and an evaluation of the materials that are involved and considered sources for M307. In the case of an drafted-only composition like M307, these will be called sketch *complexes* (analogous to work complexes), i.e., the set of individual sketches (Sk1, Sk2, up to Sk7 in Figure 9) that can be identified on one or more certain pages (fol. 3v in Figure 9) of one or more physical sign carriers (Webern's sketch book no. 3 in Figure 9). It is this level of the sketch complexes where the actual philological work takes place: the source description describes the relevant sections of the physical sign carrier, the contents of which are transcribed in the sheets. The source evaluation evaluates the physical material in terms of a source, and defines the content and form of the text-critical comments (TkA), which are themselves philological annotations to the transcribed musical texts. Again, these sub-sections of the sketch complexes (sheets, source evaluation, and source description) are back-linked, and contributing to their respective super-sections (indicated by dashed arrows). As can be seen in Figure 9, the nodes in the graph have different functionalities: some of them are digital representations of a real world entity, either physical (like the sketchbook), or abstract (like the work complex M307, or its sketch complexes). These could be referred to as the actual points of interest of an edition (marked in grey). Other nodes (marked in yellow) stand for textual manifestations (like introductions, transcribed musical texts, source lists, evaluations, or descriptions) that are solely produced by the editors within the context of their philological work. The remaining nodes (marked in cyan) represent digital *container* objects that need to be filled, either by the produced texts, or by backlinks from lower levels.

None of these philological issues is actually new or surprising compared to approaches in *traditional* editions. But in order to make philological knowledge accessible to machine-interpretable processing, it is necessary to explicitly name and model the operations from which it is shaped. One of the advantages of such graph-based modelling becomes evident in Figure 9. There is nothing to constrain the perspective from which an editor or user has to approach the content of an edition. Coming from the edition side, one could ask: Show me all sketches that are described in section II/5. Sk1 of M307 would be one of these sketches. From the material side, one could ask: Which entities in Sketchbook 3 have a source description? Here, too, Sk1 of M307 would be one of the returned items. Finally, one could merge the questions and ask: Show me all sketches that are described in section II/5 and are notated in Sketchbook 3. And again, Sk1 of M307 would be one of the results. Because of the explicit distinction between the level of the edition and the level of the (physical) material,[20] it becomes quite easy to switch perspectives without confusion or loss of orientation.

3.2 Connecting the model to the world

The proposed model is intended to be compliant with the application models of *Knora,* and to the framework of the *NIE-INE* ontologies. Besides the fact that the NIE-INE ontologies themselves are highly connected to *CIDOC CRM* and other existing ontologies, there are also various entry and connection points for the ontologies discussed earlier, in section 2 of this paper: *FRBRoo* via *DoReMus*, *MusicOWL* or CIDOC's *CRMinf.* Some of these connections are illustrated in Figure 11. According to *FRBRoo*, a physical sign carrier (the sketchbook in our example above) can be regarded as a F4 Manifestation Singleton, and the text objects produced by the editors as E31 Documents. In a way, the entity called M306 in the edition, as well as the corresponding sketch complexes associated with it, are documenting abstract works (M306 resp. first to nth sketch of M306). As already mentioned in chapter 2.1, these abstract works are represented in *DoReMus* (following *FRBRoo*) by a *F15 Complex Work* object and its corresponding member (a *F14 Individual Work*), which is realised in a *F22 Self-Contained Expression* that was created by an *F28 Expression Creation.* Following this path along the graph (marked in orange), the aforementioned digital representations of the abstract and the physical edition subjects become connected once more, this time not in terms of philological processes, but in terms of the creation processes of a musical piece. In this way, both the production processes and

[20] Peter Boot and Marijn Koolen called these two levels the *editable domain* and the *edition domain*, cf. their article in the present volume.

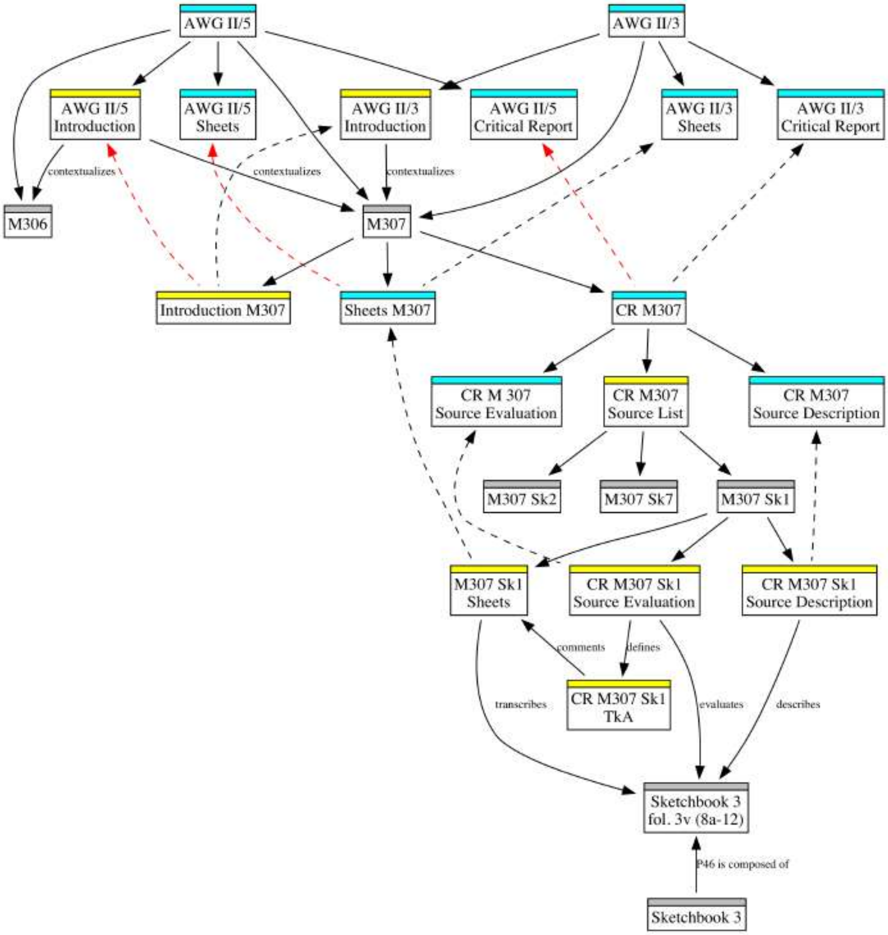


Figure 9. Philological processes concerning the sections of the edition and the material found in Webern's sketchbooks using the example of one sketch (Sk1) of M307.

reproduction processes of a piece, closely interacting with each other in an edition project, get interlinked and at the same time differentiated explicitly. Other sections of the *DoReMus* ontology including publications, performances, recordings, or reception processes can also be addressed here.

Regarding the transcribed musical texts, their status as a E31 Document can be further specified as a (digital) musical score. Utilizing the *Music Ontology* (*mo:Score*) and the *MusicOWL* ontology, this is the entry point for the highly detailed relationships on the level of music notation, as depicted in Figure 2 above. Following the approach of the *MELD* framework, these models can be used on top of a digitally encoded representation of the music score[21] to support the linking of the transcribed texts to philological observations, such as text-critical commentaries or alternative readings.

As already mentioned, the *CRMinf* Argumentation model can be particularly useful in the field of (digital) humanities. It allows scholars to deal with uncertainty, doubts, hypotheses, or any kind of argumentative conclusion. For graph-based scholarly digital editions, this model could take a genuine place in source evaluation: here, the ranking and relationships of the materials considered as sources are negotiated, their status is examined, and possible missing sources (*deperdita*) can be determined (Münnich 2019). In a final vision, however, it could be imagined that such a model could be applied to every single triple statement in the graph, in order to make the decision- and knowledge-making processes far more transparent. But this transparency would come at the price of the overall model quickly becoming much more complex. Figure 10 exemplifies a *simple* case, in which the conclusion of a scholar regarding two propositions (A & B) is accepted and adopted by a second scholar. The increasing complexity is easy to imagine if contradictions or scientific controversies and discourses are included in the modelling. But such complexity should be welcomed, as it allows scholarly argumentation in a digital context to overcome under-complexity or under-specification, which is induced by argumentatively restrictive or limited digital applications, and lags behind scientific standards and best practice.

3.3 Transforming philological knowledge

The example of some of Webern's sketches will be used to demonstrate how the idea of thinking about graphs and networks can influence the philological processes

[21] The AWG is in close contact (participation in workshops, conferences, development) and exchange with the MEI community, especially to clarify the question how a transcoding of the existing edited music texts (which are prepared with the music notation program *Finale* and available right now in MusicXML, PDF and SVG output format) into the MEI format would be possible. Since the philological findings and procedures require extensive manual intervention and adjustments to the *Finale* transcriptions, such transcoding can only be carried out to a certain extent (semi-)automatically. Until a technically feasible solution can be found within the capacities of the AWG in the medium term, the score texts will be embedded as SVG graphics within the online edition of the AWG.

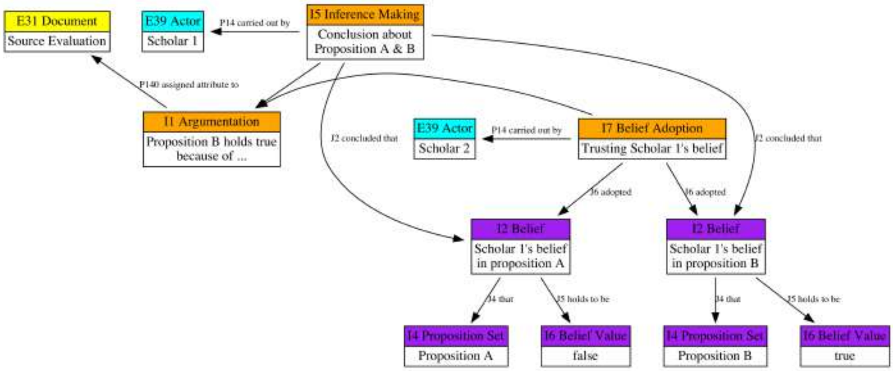


Figure 10. Example of an argumentative conclusion assigned to a source evaluation.

and the work of an edition project. Figure 12 shows a detail of folio 3^v in Webern's Sketchbook 3 on which the first sketches for *Der Spiegel sagt mir: ich bin schön!* M307 can be found.

One full page width sketch (Sk1) is accompanied by two smaller sketches above it (Sk1.1 and Sk1.2). From a philological examination, it becomes clear that Sk1.1 must have been created in parallel to the main sketch Sk1, since both sketches reflect and influence the changes of each other, without it being possible to determine which changes came first. We would call this a concomitant, accompanying relationship between the two sketches. In contrast, Sk1.2 must obviously have been created after finishing Sk1.1, since its first layer includes changes made in Sk1.1. This is what we would call a preceding, consecutive relationship. (It should be noted that Sk1.2 and Sk1 are in turn in a concomitant relationship to each other, i.e., here both sketches reflect and influence changes in the other.)

Going through the entire three pages in Sketchbook 3 that are related to M307, different working stages can be identified (Figure 13). They clearly start with an arrangement for four voices, then reduce the casting to three voices, before finally experimenting with the casting for one voice and piano, so transforming into a piano song, instead of a purely vocal composition. Additionally, different variants of the underlying twelve tone row can be found, as well as some paratexts that are connected to the start and end dates of the compositional process (between July 7^{th} and 9^{th} 1930).

Allocating all (up to 13) of the larger and smaller sketches of M307 (here named Sk1 to Sk8) to these different stages results, almost automatically, in a graph-based visualization and orientation of the dependencies and interrelations between the

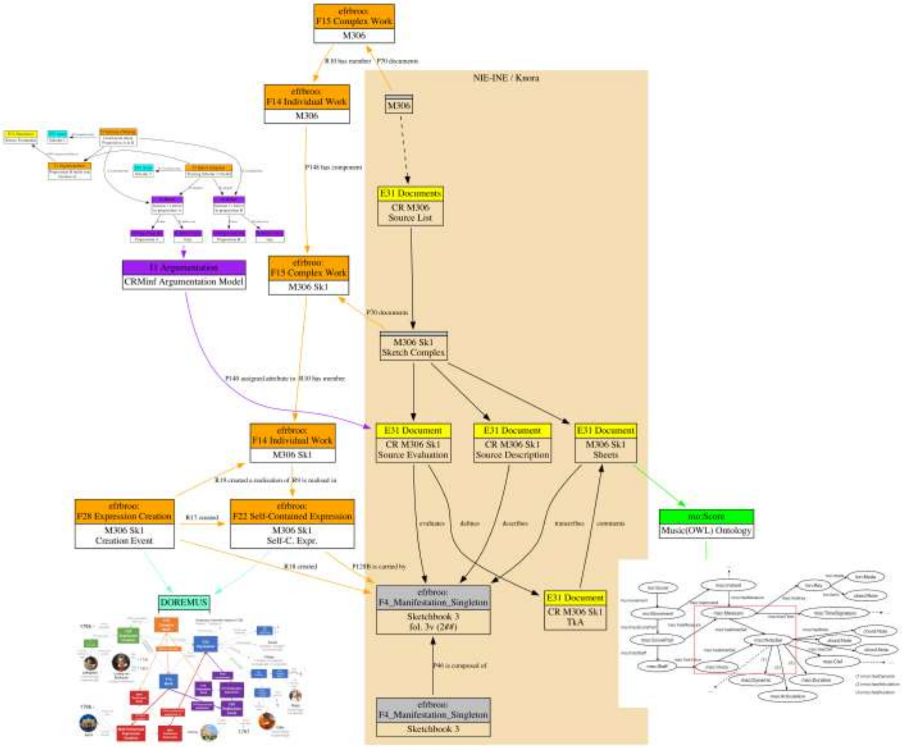


Figure 11. AWG embedded in the context of existing ontologies like *FRBRoo*, *DoReMus*, *MusicOWL*, or *CRMinf*.

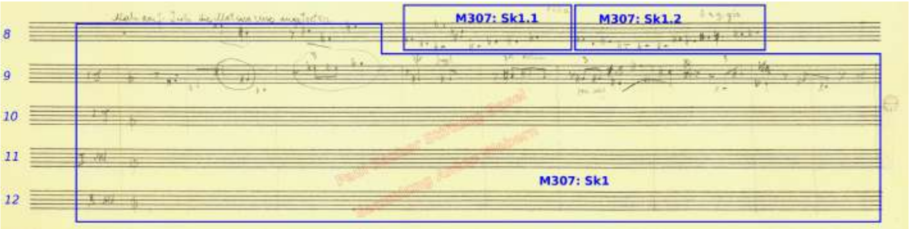


Figure 12. Folio 3^v of Webern's Sketchbook 3 with first sketches of M307 (detail of staves 8–12).

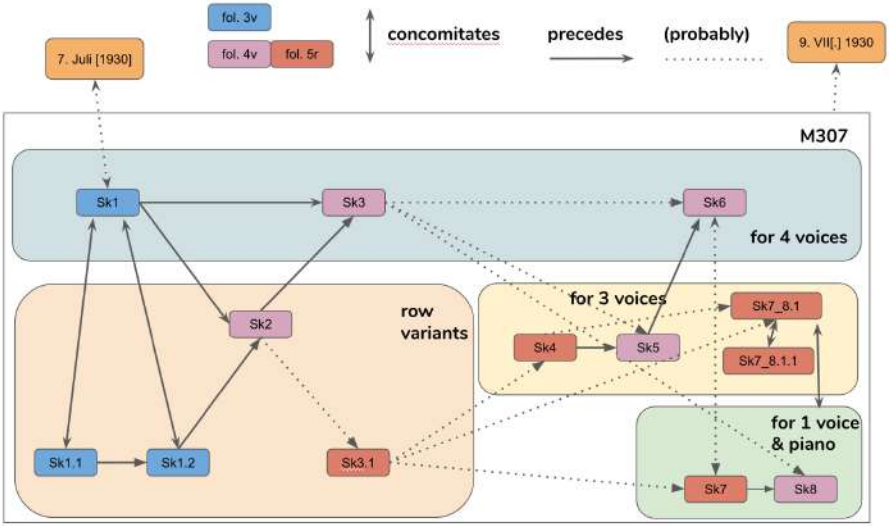


Figure 13. Philological *pathways* through Webern's M307. © Anton Webern Gesamtausgabe, 2019. CC-BY-NC-SA 4.0.

sketch complexes, as shown in Figure 13. Hereby, every established relationship (indicated by arrows in the figure) provides a short evaluation and explanation as to how the conclusion was reached by the editors. Expressed in RDF triples, all of this combines to form a graph-based source evaluation that allows for multiple *pathways*[22] through the different source materials involved. While the editors can provide their particular view and their particular *pathway*, other researchers and users could follow another path.

4 Conclusion

A quarter of a century after the first propositions and attempts to build a (then called) computer edition and its critical apparatus as an "hypertext presentation in the World Wide Web" (Peter & Wender 1997; Hoffmann et al. 1993), scholarly digital editions have adapted to the new possibilities and challenges of the technological and conceptual developments of the web. Among various approaches, XML-based solutions have played a crucial role for a long time. XML gains its full power in connection with

[22] The concept of *pathways* for digital (music) editions was recently proposed and discussed by Kepper and Pugin 2017, 362–363.

encodings of a document's structure. But it has its structural, conceptual and semantic limitations, like any artificial model, including RDF. But RDF, as with other graph-based approaches, provides another complementary perspective, and adds a level of differentiation that goes beyond the expressiveness of XML.

In this paper, we have tried to give another example of the potential of RDF-based modelling of philological knowledge. The possibilities for the representation of semantic relationships and the links between different areas of a scholarly digital (music) edition, the overlap-free representation of non-hierarchical content, as well as the connection, interlinking and interoperability with external data sets, appear to be the main advantages of such an approach. Of course, these structural possibilities could also be applied and continued beyond the scope of our example: thus, the indicated *pathways* through the sketches of Webern's M307 could be extended to paths through Webern's entire oeuvre, which then can be examined from within its historical context; ultimately, the presentation of a comprehensive music history would also be conceivable, progressing from the smallest surviving source materials, to larger musical or cultural perspectives.

Finally, we have to take into consideration that a graph-based, semantic approach to scholarly (music) editions is not only about using cutting-edge technology, it is about transforming philological knowledge into a machine-interpretable environment, and about changing the way in which we ourselves are enabled to think about philological and music historical processes.

5 Future Work

The proposed model for scholarly music editions is a work in progress, and a lot needs to be done: the classes and concepts to be re-used from external ontologies must be finally determined and applied to the model, especially those from the *Music Ontology*, *MusicOWL*, or from *DoReMus*. To take this step, the interlinking and interaction of the Knora models, the *NIE-INE* ontologies, and the *DoReMus* ontologies have to be further tested and investigated. On the level of music notation encoding, it would be great to see a closer connection between MEI format and *MusicOWL*. Hereby, the fundamental research and work of the *MELD* project will be of great assistance.

Acknowledgements

We are very grateful to Hans Cools, David M. Weigl, Iian Neill, Terhi Nurmikko-Fuller, Dominic Oldman, Roberta Padlina, Kevin R. Page, John Pybus, Stephen Stead, and Barbara Wiermann, who strongly inspired our thoughts on this subject, and especially Jim Jones and Albert Meroño-Peñuela for their helpful suggestions and comments on the draft of this paper. The Anton Webern Gesamtausgabe is funded by the Swiss National Science Foundation (SNSF no. 157968) and the Swiss Academy of Humanities and Social Sciences.

Bibliography

Achichi, Manel, Rodolphe Bailly, Cécile Cecconi, Marie Destandau, Konstantin Todorov, and Raphaël Troncy, 'DOREMUS: Doing Reusable Musical Data', in *Proceedings of the ISWC 2015 Posters & Demonstrations Track, Bethlehem, PA, USA, October 11, 2015*, ed. by Serena Villata, Jeff Z. Pan, and Mauro Dragoni, CEUR Workshop Proceedings 1486, 2015 <http://ceur-ws.org/Vol-1486/paper_75.pdf>

Acquavella-Rauch, Stefanie, '(Musik)Edition im "digitalen" Zeitalter – Versuch einer Verortung konzeptioneller und struktureller Veränderungen', in *Beitragsarchiv des Internationalen Kongresses der Gesellschaft für Musikforschung, Mainz 2016 – "Wege der Musikwissenschaft"*, ed. by Gabriele Buschmeier and Klaus Pietschmann (Mainz: Schott Campus, 2019) <urn:nbn:de:101:1-2019020610345665991353>

AWG, 'Edition prototype of the Anton Webern Gesamtausgabe', 2015a <https://edition.anton-webern.ch/>

———, 'Project website of the Anton Webern Gesamtausgabe', 2015b <https://www.anton-webern.ch/>

Bangert, Daniel, Terhi Nurmikko-Fuller, and Alfie Abdul-Rahman, 'JazzCats Project', 2016 <http://jazzcats.cdhr.anu.edu.au/>

Bangert, Daniel, Terhi Nurmikko-Fuller, J. Stephen Downie, and Yun Hao, 'Jazzcats: Navigating an RDF Triplestore of Integrated Performance Metadata', in *Proceedings of the 5th International Conference on Digital Libraries for Musicology (DLfM '18), Paris, France, September 28, 2018* (presented at the DLfM '18, New York, NY: ACM Press, 2018), 74–77 <https://doi.org/10.1145/3273024.3273031>

Beethovens Werkstatt, 'Project website', 2014 <https://beethovens-werkstatt.de/>

Belhajjame, Khalid, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and others, *PROV-O: The PROV Ontology*, 30 April 2013 <http://www.w3.org/TR/2013/REC-prov-o-20130430/>

Berners-Lee, Tim, 'What the Semantic Web Can Represent', 1998 <https://www.w3.org/DesignIssues/RDFnot.html>

Berners-Lee, Tim, James Hendler, and Ora Lassila, 'The Semantic Web', *Scientific American*, 284.5 (2001), 34–43 <https://www.jstor.org/stable/26059207>

Bleier, Roman, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider, eds., *Digital Scholarly Editions as Interfaces*, Schriften des Instituts für Dokumentologie und Editorik, 12 (Norderstedt: Books on Demand, 2018) <urn:nbn:de:hbz:38-90853>

Brickley, Dan, and Libby Miller, *FOAF Vocabulary Specification*, 14 January 2014 <http://xmlns.com/foaf/spec/20140114.html>

Choffé, Pierre, and Françoise Leresche, 'DOREMUS : Connecting Sources, Enriching Catalogues and User Experience', 2016 <http://library.ifla.org/id/eprint/1322>

CIDOC, 'CIDOC Documentation Standards Working Group, and CIDOC CRM SIG', 2006 <http://www.cidoc-crm.org/>

Corthaut, Nik, Sten Govaerts, Verbert Katrien, and Erik Duval, 'Connecting the Dots: Music Metadata Generation, Schemas and Application', in *Proceedings of the 9th International Conference of Music Information Retrieval (ISMIR 2008)*, ed. by Juan Pablo Bello, Elaine Chew, and Douglas Turnbull (presented at the ISMIR 2008, Philadelphia, PA, USA, 2008), 249–254 <http://ismir2008.ismir.net/papers/ISMIR2008_213.pdf>

Daquino, Marilena, Enrico Daga, Mathieu D'Aquin, Aldo Gangemi, Simon Holland, Robin Laney, and others, 'Characterizing the Landscape of Musical Data on the Web: State of the Art and Challenges', 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*, CEUR Workshop Proceedings 2014 (presented at the WHiSe 2017, Vienna, 2017), 57–68 <http://ceur-ws.org/Vol-2014/paper-07.pdf>

Davis, Randall, Howard Shrobe, and Peter Szolovits, 'What Is a Knowledge Representation?', *AI Magazine*, 14.1 (1993), 17–33 <https://groups.csail.mit.edu/medg/ftp/psz/k-rep.html>

DCMI, *Dublin Core Metadata Initiative Schemas*, 1995 <https://www.dublincore.org/schemas/>

Digital Mozart Edition (DME), 'Project website', 2006 <https://dme.mozarteum.at/>

Doerr, Martin, Stephen Stead, and Paveprime Ltd., *CRMinf: The Argumentation Model*, February 2015 <http://www.cidoc-crm.org/crminf/ModelVersion/version-0.7>

Driscoll, Matthew James, and Elena Pierazzo, eds., *Digital Scholarly Editing. Theories and Practices* (Cambridge, UK: Open Book Publishers, 2016) <https://doi.org/10.11647/OBP.0095>

EDM, 'Documentation of the Europeana Data Model', 2012 <https://pro.europeana.eu/page/edm-documentation>

Emans, Reinmar, and Ulrich Krämer, eds., *Musikeditionen im Wandel der Geschichte*, Bausteine zur Geschichte der Edition, 5 (Berlin, München, Boston: De Gruyter, 2015) <https://doi.org/10.1515/9783110434354>

Freischütz Digital, 'Project website', 2012 <https://www.freischuetz-digital.de/>

Gerhard, Anselm, ed., *Musikwissenschaft – eine verspätete Disziplin? Die akademische Musikforschung zwischen Fortschrittsglauben und Modernitätsverweigerung* (Stuttgart: J.B. Metzler, 2000)

Goebl, Werner, and David M. Weigl, 'Digitising and Enriching Our Cultural Heritage Together / Die Digitalisierung und Anreicherung unseres musikalischen Erbes selbst gestalten', *mdw-Webmagazin*, 27 September 2019 <https://web.archive.org/web/20200504064753/https://www.mdw.ac.at/magazin/index.php/2019/09/27/die-digitalisierung-und-anreicherung-unseres-musikalischen-erbes-selbst-gestalten/>

Good, Michael, *MusicXML*, 7 December 2017 <https://web.archive.org/web/20190930143620/https://w3c.github.io/musicxml/>

———, 'MusicXML for Notation and Analysis', in *The Virtual Score: Representation, Retrieval, Restoration*, ed. by Walter B. Hewlett and Eleanor Selfridge-Field, Computing in Musicology, 12 (Cambridge, MA; Stanford, CA: MIT Press ; CCARH, Stanford University, 2001), 113–124

Harley, Nicholas, and Geraint Wiggins, 'An Ontology for Abstract, Hierarchical Music Representation', in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR 2015). Málaga, Spain*, 2015

Hoffmann, Dirk, Peter Jörgensen, and Otmar Foelsche, 'Computer-Edition statt Buch-Edition. Notizen zu einer historisch-kritischen Edition - basierend auf dem Konzept von Hypertext und Hypermedia', *Editio*, 7 (1993), 211–220 <https://doi.org/10.1515/9783110241983.211>

IFLA Study Group, *Functional Requirements for Bibliographic Records*, IFLA Series on Bibliographic Control 19 (Munich, 1998) <https://web.archive.org/save/https://www.ifla.org/publications/functional-requirements-for-bibliographic-records>

IFLA Working Group on FRBR/CRM Dialogue, *Definition of FRBRoo. A Conceptual Model for Bibliographic Information in Object-Oriented Formalism*, ed. by Chryssoula Bekiari, Martin Doerr, Patrick Le Bœuf, and Pat Riva, Version 2.4, 2015 <https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf>

International Image Interoperability Framework Consortium, *Specifications of the International Image Interoperability Framework*, 2019 <https://iiif.io/technical-details/#stable-specifications>

Jones, Jim, Kleber Tertuliano, Diego de Siqueira Braga, and Tomi Kauppinen, *MusicOWL - Music Score Ontology*, 6 July 2017a <http://linkeddata.uni-muenster.de/ontology/musicscore#1.0.0>

———, 'MusicOWL. The Music Score Ontology', in *Proceedings of the International Conference on Web Intelligence – WI '17* (presented at the WI '17, New York, NY: ACM Press, 2017b), 1222–1229 <https://doi.org/10.1145/3106426.3110325>

Kallionpää, Maria, Chris Greenhalgh, Adrian Hazzard, David M. Weigl, Kevin R. Page, and Steve Benford, 'Composing and Realising a Game-Like Performance for Disklavier and Electronics', in *NIME 2017. New Interfaces for Musical Expression, Copenhagen, 15-18 May 2017. Proceedings*, ed. by Cumhur Erkut, 2017, 464–469 <http://eprints.nottingham.ac.uk/id/eprint/44529>

Kamzelak, Roland S., 'Digitale Editionen im Semantic Web. Chancen und Grenzen von Normdaten, FRBR und RDF', in *„Ei, Dem alten Herrn zoll' ich Achtung gern'". Festschrift für Joachim Veit zum 60. Geburtstag*, ed. by Kristina Richts and Peter Stadler (München: Allitera Verlag, 2016), 423–36 <https://doi.org/10.25366/2018.29>

Kepper, Johannes, *Musikedition im Zeichen neuer Medien. Historische Entwicklung und gegenwärtige Perspektiven musikalischer Gesamtausgaben*, Schriften des Instituts für Dokumentologie und Editorik, 5 (Norderstedt: Books on Demand, 2011) <https://kups.ub.uni-koeln.de/6639/>

———, 'XML-basierte Codierung musikwissenschaftlicher Daten — Zu den Voraussetzungen einer digitalen Musikedition // XML-based Encoding of Musicological Data — About the Requirements of a Digital Music Philology', *it – Information Technology. Methoden und innovative Anwendungen der Informatik und Informationstechnik*, 51.4 (2009), 216–221 <https://doi.org/10.1524/itit.2009.0544>

Kepper, Johannes, and Laurent Pugin, 'Was ist eine Digitale Edition? Versuch einer Positionsbestimmung zum Stand der Musikphilologie im Jahr 2017', *MusikTheorie. Zeitschrift für Musikwissenschaft*, 32.4 (2017), 347–363

Klumpenhouwer, Henry, ‘Network Analysis and Webern’s Opus 27/III’, *Tijdschrift Voor Muziektheorie*, 3.1 (1998), 24–37

Lewin, David, *Generalized Musical Intervals and Transformations* (New Haven: Yale University Press, 1987)

———, ‘Klumpenhouwer Networks and Some Isographies That Involve Them’, *Music Theory Spectrum*, 12.1 (1990), 83–120 <https://doi.org/10.2307/746147>

Lisena, Pasquale, Raphaël Troncy, Konstantin Todorov, and Manel Achichi, ‘Modeling the Complexity of Music Metadata in Semantic Graphs for Exploration and Discovery’, in *Proceedings of the 4th International Workshop on Digital Libraries for Musicology (DLfM’17), Shanghai, China, October 28, 2017* (presented at the DLfM ’17, New York, NY: ACM Press, 2017), 17–24 <https://doi.org/10.1145/3144749.3144754>

Mazzola, Guerino, *Geometrie der Töne. Elemente der Mathematischen Musiktheorie* (Basel: Birkhäuser Basel, 1990) <https://doi.org/10.1007/978-3-0348-7427-4>

———, *The Topos of Music. Geometric Logic of Concepts, Theory, and Performance* (Basel: Birkhäuser Basel, 2002) <https://doi.org/10.1007/978-3-0348-8141-8>

Mazzola, Guerino, and Moreno Andreatta, ‘Diagrams, Gestures and Formulae in Music’, *Journal of Mathematics and Music*, 1.1 (2007), 23–46 <https://doi.org/10.1080/17459730601137716>

Meroño-Peñuela, Albert, and Rinke Hoekstra, *MIDI Ontology*, 2016a <http://purl.org/midi-ld/midi#>

———, ‘The Song Remains the Same: Lossless Conversion and Streaming of MIDI to RDF and Back’, in *The Semantic Web – ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29 – June 2, 2016, Revised Selected Papers*, ed. by Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer, and Christoph Lange, Lecture Notes in Computer Science 9989 (presented at the ESWC 2016, Cham: Springer International Publishing, 2016b), 194–199 <https://doi.org/10.1007/978-3-319-47602-5_38>

MIDI Manufacturers Association (MMA), *The Official MIDI Specifications*, 2019 <https://web.archive.org/web/20190930145741/https://www.midi.org/specifications>

Miles, Alistair, and Sean Bechhofer, *SKOS Simple Knowledge Organization System Reference*, 18 August 2009 <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>

Moldenhauer, Hans, and Rosaleen Moldenhauer, *Anton von Webern. A Chronicle of His Life and Work* (London: Victor Gollancz Ltd., 1978)

motools, ‘Motools [Github Repo]’, 2007 <https://github.com/motools>

———, ‘Project Website Music Ontology’, 2013 <http://musicontology.com/>

Münnich, Stefan, ‘Ontologien als semantische Zündstufe für die digitale Musikwissenschaft? // Ontologies as a Semantic Booster for Digital Musicology? An Overview’, *Bibliothek. Forschung und Praxis*, 42.2 (2018), 184–193 <https://doi.org/10.1515/bfp-2018-0027>

———, ‘Quellenverluste (Deperdita) als methodologischer Unsicherheitsbereich für Editorik und Datenmodellierung am Beispiel von Anton Weberns George-Lied op. 4 Nr. 5’, in *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, ed. by Andreas Kuczera, Thorsten Wübbena, and Thomas Kollatz, Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4, text/html format (Wolfenbüttel, 2019) <https://doi.org/10.17175/sb004_005>

Music Encoding Initiative, *Guidelines*, 2018

———, ‘Project Website’, 2017 <https://music-encoding.org/>

MusicML, *Musicml.Owl*, 10 September 2016 <https://web.archive.org/web/20190930143748/http://www.ontologydesignpatterns.org/ont/musicml/musicml.owl>

Nurmikko-Fuller, Terhi, and Kevin R. Page, ‘A Linked Research Network That Is Transforming Musicology’, in *Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe I) Co-Located with 13th ESWC Conference 2016 (ESWC 2016). Anissaras, Greece, May 20th, 2016*, ed. by Alessandro Adamou, Enrico Daga, and Leif Isaksen, CEUR Workshop Proceedings 1608, 2016, 73–78 <http://ceur-ws.org/Vol-1608/paper-09.pdf>

Oldman, Dominic, Martin Doerr, and Stefan Gradmann, ‘Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge’, in *A New Companion to Digital Humanities*, ed. by Susan Schreibman, Ray Siemens, and John Unsworth (Chichester, UK: John Wiley & Sons, Ltd, 2016), 251–273 <https://doi.org/10.1002/9781118680605.ch18>

Oldman, Dominic, and Diana Tanase, ‘Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace’, in *The Semantic Web – ISWC 2018. 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II*, ed. by Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, and others, Lecture Notes in Computer Science 11137 (presented at the ISWC 2018, Cham: Springer International Publishing, 2018), 325–340 <https://doi.org/10.1007/978-3-030-00668-6_20>

Peter, Robert, and Herbert Wender, ‘Variantenapparate als Hypertext im Internet. Perspektiven einer Computer-Edition’, in *Textproduktion in elektronischen Umgebungen*, ed. by Dagmar Knorr and Eva-Maria Jakobs, Textproduktion und Medien (Frankfurt a.M.: Peter Lang, 1997), II, 141–154 <http://www.prowitec.rwth-aachen.de/p-publikationen/band-pdf/band2/band2_peter_wender.pdf>

Pierazzo, Elena, *Digital Scholarly Editing. Theories, Models and Methods* (Farnham, Surrey, UK; Burlington, VT: Ashgate, 2015)

Popoff, Alexandre, Moreno Andreatta, and Andrée Ehresmann, ‘Relational Poly-Klumpenhouwer Networks for Transformational and Voice-Leading Analysis’, *Journal of Mathematics and Music*, 12.1 (2018), 35–55 <https://doi.org/10.1080/17459737.2017.1406011>

Raimond, Yves, Samer Abdallah, Mark Sandler, and Frederick Giasson, ‘The Music Ontology’, in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), Vienna, Austria, September 23-27, 2007*, ed. by Simon Dixon, David Bainbridge, and Rainer Typke (presented at the ISMIR 2007, Wien: Austrian Computer Society, 2007), 417–422 <http://ismir2007.ismir.net/proceedings/ISMIR2007_p417_raimond.pdf>

Raimond, Yves, Thomas Gängler, Frédérick Giasson, Kurt Jacobson, George Fazekas, Simon Reinhardt, and others, *The Music Ontology Specification*, 22 July 2013 <https://web.archive.org/web/20190930143118/http://musicontology.com/specification/>

RDF Working Group, ‘Resource Description Framework (RDF)’, *W3C Semantic Web*, 2014 <https://www.w3.org/2001/sw/wiki/index.php?title=RDF&oldid=4387>

Rings, Steven, *Tonality and Transformation*, Oxford Studies in Music Theory (New York, NY: Oxford University Press, 2011)

Sahle, Patrick, *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*, Schriften des Instituts für Dokumentologie und Editorik, 7–9 (Norderstedt: Books on Demand, 2013) <https://www.i-d-e.de/publikationen/schriften/s7-9-digitale-editionsformen/>

Sanderson, Robert, Paolo Ciccarese, and Benjamin Young, *Web Annotation Vocabulary*, 23 February 2017 <https://www.w3.org/TR/2017/REC-annotation-vocab-20170223/>

Sandler, Mark, Yves Raimond, and Christopher Sutton, 'Interlinking Music-Related Data on the Web', *IEEE MultiMedia*, 16.2 (2009), 52–63 <https://doi.org/10.1109/MMUL.2009.29>

Schweizer, Tobias, and Lukas Rosenthaler, 'Building Digital Editions on the Basis of a Virtual Research Environment', in *Proceedings of the Digital Humanities Congress 2012*, ed. by Clare Mills, Michael Pidd, and Esther Ward, Studies in the Digital Humanities (presented at the DHC 2012, Sheffield: The Digital Humanities Institute, 2014) <https://www.dhi.ac.uk/openbook/chapter/dhc2012-schweizer>

Simonetta, Federico, 'Graph Based Representation of the Music Symbolic Level. A Music Information Retrieval Application' (unpublished Master's Thesis, Università di Padova, 2018) <https://doi.org/10.5281/zenodo.1476564>

Smaill, Alan, Geraint Wiggins, and Mitch Harris, 'Hierarchical Music Representation for Composition and Analysis', *Computers and the Humanities*, 27.1 (1993), 7–17 <https://doi.org/10.1007/BF01830712>

Stachowiak, Herbert, *Allgemeine Modelltheorie* (Wien: Springer, 1973)

Stead, Stephen, 'CRMinf: The Argumentation Model' (unpublished slides presented at the CIDOC CRM:SIG meeting, Oxford UK, 2015) <http://slideplayer.com/slide/9773467/>

Studer, Rudi, Richard Benjamins, and Dieter Fensel, 'Knowledge Engineering: Principles and Methods', *Data & Knowledge Engineering*, 25.1–2 (1998), 161–197 <https://doi.org/10.1016/S0169-023X(97)00056-6>

Swartz, Aaron, 'MusicBrainz: A Semantic Web Service', *IEEE Intelligent Systems*, 17.1 (2002), 76–77 <https://doi.org/10.1109/5254.988466>

Szeto, Wai Man, and Man Hon Wong, 'A Graph-Theoretical Approach for Pattern Matching in Post-Tonal Music Analysis', *Journal of New Music Research*, 35.4 (2006), 307–321 <https://doi.org/10.1080/09298210701535749>

Tymoczko, Dmitri, *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*, Oxford Studies in Music Theory (New York: Oxford University Press, 2011)

Urbanek, Nikolaus, 'Was ist eine musikphilologische Frage?', in *Historische Musikwissenschaft. Grundlagen und Perspektiven*, ed. by Michele Calella and Nikolaus Urbanek (Stuttgart: J.B. Metzler, 2013), 147–183 <https://doi.org/10.1007/978-3-476-05348-0_8>

Veit, Joachim, 'Es bleibt nichts, wie es war – Wechselwirkungen zwischen digitalen und "analogen" Editionen, *Editio*, 24.1 (2010), 37–52 <https://doi.org/10.1515/9783110223163.0.37>

———, 'Musikedition 2.0: Das "Aus" für den Edierten Notentext?', *Editio*, 29.1 (2015), 70–84 <https://doi.org/10.1515/editio-2015-006>

———, 'Musikwissenschaft und Computerphilologie – eine schwierige Liaison?', in *Jahrbuch für Computerphilologie 7 (2005)*, ed. by Georg Braungart, Peter Gendolla, and Fotis Jannidis

(Paderborn: mentis, 2006), 67–92 <http://computerphilologie.digital-humanities.de/jg05/veit.html>

Viglianti, Raffaele, 'The Music Addressability API: A Draft Specification for Addressing Portions of Music Notation on the Web', in *Proceedings of the 3rd International Workshop on Digital Libraries for Musicology*, DLfM 2016 (New York, USA: Association for Computing Machinery, 2016), 57–60 <https://doi.org/10.1145/2970044.2970056>

W3C Schema.org Community Group, *schema.org*, 1 August 2019 <https://github.com/schemaorg/schemaorg/blob/main/data/releases/3.9/schema-all.html>

Walton, Adrian, 'A Graph Theoretic Approach to Tonal Modulation', *Journal of Mathematics and Music*, 4.1 (2010), 45–56 <https://doi.org/10.1080/17459730903370940>

Webern, Anton, *Sketches (1926–1945)*, ed. by Hans Moldenhauer (New York, NY: Carl Fischer, 1967)

Weigl, David M., Werner Goebl, Tim Crawford, Aggelos Gkiokas, Nicolas F. Gutierrez, Alastair Porter, and others, 'Interweaving and Enriching Digital Music Collections for Scholarship, Performance, and Enjoyment', in *6th International Conference on Digital Libraries for Musicology*, DLfM '19 (The Hague, Netherlands: Association for Computing Machinery, 2019), 84–88 <https://doi.org/10.1145/3358664.3358666>

Weigl, David M., and Kevin R. Page, 'A Framework for Distributed Semantic Annotation of Musical Score: "Take It to the Bridge!"', in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), Suzhou, China, October 23-27, 2017*, ed. by Xiao Hu, Sally Jo Cunningham, Douglas Turnbull, and Zhiyao Duan (presented at the ISMIR 2017, Suzhou, China, 2017), 221–228 <https://ismir2017.smcnus.org/wp-content/uploads/2017/10/190_Paper.pdf>

Wettlaufer, Jörg, 'Der nächste Schritt? Semantic Web und digitale Editionen', in *Digitale Metamorphose: Digital Humanities und Editionswissenschaft*, ed. by Roland S. Kamzelak and Timo Steyer, Zeitschrift für digitale Geisteswissenschaften / Sonderbände 2, text/html format (Wolfenbüttel, 2018) <https://doi.org/10.17175/sb002_007>

Wiggins, Geraint A., Mitch Harris, and Alan Smaill, 'Representing Music for Analysis and Composition', in *Proceedings of the Second Workshop on AI and Music*, ed. by M. Balaban, K. Ebcioglu, O. Laske, C. Lischka, and L. Soriso (Menlo Park, CA: AAAI, 1989), 63–71

Modelling Cross-Document Interdependencies in Medieval Charters of the St. Katharinenspital in Regensburg

Colin Sippl, Manuel Burghardt, Christian Wolff

Abstract

To overcome the limitations of structural XML mark-up, *graph-based data models* and graph databases, as well as *event-based ontologies* like *CIDOC-CRM* (FORTH-ICS 2018) have been considered for the creation of *digital editions*. We apply the graph-based approach to model charter regests and extend it with the CIDOC-CRM ontology, as it allows us to integrate information from different sources into a flexible data model. By implementing the ontology within the *Neo4j* graph database (Neo4j 2018) we create a sustainable data source that allows *explorative search queries* and finally, the integration of the database in various technical systems. Our use case are the charters from the *St. Katharinenspital*, a former medieval hospital in Regensburg, Germany. By analysing charter abstracts with *natural language processing (NLP)* methods and using additional data sources related to the charters, we generate additional metadata. The extracted information allows the *modelling of cross-document interdependencies of charter regests* and their related entities. Building upon this, we develop an *exploratory web application* that allows to investigate a graph-based digital edition. Thereby, each entity is displayed in its unique context, i.e., it is shown together with its related entities (next neighbours) in the graph. We use this to enhance the result lists of a full-text search, and to generate entity-specific detail pages.

1 Introduction

The creation of digital editions is one of the main applications in the Digital Humanities. In recent years, the use of the eXtensible Markup Language (XML) along with Text Encoding Initiative (TEI) styles has become the most common way to do this (Sahle 2013, 341–42). However, Kuczera (2016) and Sahle (2013, 345–71) point out some severe limitations of digital editions that merely rely on structural XML mark-up. Among the main drawbacks of XML is its inability to model overlapping structures or parallel annotation hierarchies (Kuczera 2016). To overcome these limitations, graph-based data models and graph databases (Kuczera 2017), as well as event-based ontologies like *CIDOC-CRM* (Le Bœuf et al. 2015), have been considered for the creation of digital

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 181–203.

editions (Ore 2009). The graph-based approach has already been tested with charters of the *Monasterium.Net*[1] platform (Jeller 2019). However, existing implementations are highly experimental and only highlight basic aspects of the graph-based approach, e.g., the ability to analyse mutual relationships as well as new visualisation types. Finally, these use cases are unsuitable for end users, as they are just proof-of-concepts. We apply the graph-based approach of modelling digital editions and extend it with the CIDOC-CRM ontology, as it allows us to integrate information from different sources into a flexible data model. We achieve this by preparing raw data in such a way that charters from our data source can easily be linked together in a graph database. By using the CIDOC-CRM ontology, we create a sustainable data source that allows for specific and explorative search queries and, finally, integration into various technical systems. For this purpose, we rely on the *Neo4j graph database.*[2] Thereby we are shifting the focus from the structure of the documents to their *entities* and *relations.* Our use case are charters from the *St. Katharinenspital*, a former medieval hospital in Regensburg, Germany. These charters bear witness to changes in the power structure in Regensburg over the course of several centuries (Kaufner 2011, 13–21). In the following, we give an overview of our dataset, the process of data preparation, data analysis, as well as our data model. Building upon this, we develop an exploratory web application[3] that can be used to investigate a graph-based digital edition of charters from the St. Katharinenspital. Our implementation serves as a blueprint for the future development of a digital repository for graph-based scholarly editions of medieval charters.

2 The St. Katharinenspital Dataset

In total, the archive of the St. Katharinenspital holds more than 4,000 charters, which is a considerable number for a rather small archive (König 2003, 16). Hence, the archive contains an essential part of the historical written heritage of Regensburg and its surrounding area. The dataset from which the entities and relations are extracted consists of three main components: the *CEI-XML*[4] charter regests on Monasterium.Net, the scholarly editions series *Die älteren Urkunden des St. Katharinenspitals in Regensburg* and other works,[5] as well as a dataset of different files from the archives of

[1] St. Katharinenspital dataset on Monasterium.Net, available at http://monasterium.net/mom/DE-AKR/Urkunden/fond, all hyperlinks in this article were last accessed on Oct. 7^{th}, 2019.

[2] Neo4j Graph Database (Neo4j 2018).

[3] Currently only available in German language, available at https://urkunden.ur.de.

[4] Charters Encoding Initiative, CEI – The Project. Mark-up for medieval and early modern legal records (Vogeler 2004).

[5] *Die älteren Urkunden des St. Katharinenspitals in Regensburg* currently consists of three printed volumes (König 2003; Kaufner 2011; Sturm 2013). Additionally, we incorporate an unpublished volume (Feichtmeier, n.d.) and an earlier work about medieval chancery in Regensburg (Ambronn 1968).

St. Katharinenspital, including MS Office files, scans and digital documents containing the historical tradition of the St. Katharinenspital charters. Currently, 1,050 charter regests are already available on Monasterium.Net as CEI-XML documents. These documents are of varying lengths, levels of detail, and represent different states of primary source analysis. The documents contain 712 places (292 distinct) and 1,217 (749 distinct) persons as manually annotated entities.[6] Due to missing links between the CEI-XML documents and a lack of standardisation, uniform spellings of those entities are largely missing. Furthermore, a mixture of entity names with additional (e.g., biographical) data frequently[7] occurs:

```
<cei:back>
  <cei:persName>Karl der Große, fränkischer König und Kaiser</cei:persName>
  <cei:placeName>Frankfurt a. Main (krfr.St., Hessen)</cei:placeName>
  <cei:placeName>Vivarias (Gewässer bei Regensburg)</cei:placeName>
  <cei:placeName>Pielmühle (Gde. Lappersdorf, Lkr. Regensburg)</cei:placeName>
</cei:back>
```
[8]

The example shown above highlights a structural problem of documents annotated with flexible data schemes such as CEI-XML. Since common goals for advanced annotation and analysis of primary documents can be achieved in different ways, ambiguities may occur among the individual documents. The exact spelling of place names or person names (e.g., “Karl der Große”), along with any normalisation efforts, are ultimately left to the editor of a CEI-XML file on Monasterium.Net (Jeller 2019). Because of these varying spellings and structural differences between single CEI-XML files in the dataset[9], linking and analysing the documents is severely impeded. Additionally, for the St. Katharinenspital dataset there are no finding aides, like indices or registers, online. Moreover, archive IDs are inconsistently distributed among our Monasterium.Net dataset, and personal identifier (PID) references to authority files like the *Gemeinsame Normdatei* (Integrated Authority File, GND: Deutsche Nationalbibliothek 2018) are missing. Together with the limited full text search capabilities of Monasterium.Net, the issues related to the St. Katharinenspital dataset pose restrictions to working with the CEI-XML data as well as linking them to external data sources. Besides, an additional 1,699 regests (as of 2018) have been transcribed or extended by means of Microsoft Office tools (Word, Excel and Access) and are part of the data collection of the St. Katharinenspital. The complete transcription of *Repertorium C*, known as *DicMihi*, is one of the most important file sources. The original source is a register from 1745, which lists the St. Katharinenspital charters

[6] These numbers were determined by XQuery queries and after an extensive data cleansing process.

[7] E.g., “Otto Prager (1243–1244, 1248–1251, 1255)” or “Eberhard, Graf von Abensberg, Erzdiakon”.

[8] München, Bayerisches Hauptstaatsarchiv Kloster St. Emmeram Regensburg Urkunden (0794-1800) BayHStA, Kloster St. Emmeram Regensburg Urkunden 1, available at http://monasterium.net/mom/DE-BayHStA/KURegensburgStEmmeram/000001/charter.

[9] I.e., inconsistent usage of XML tags.

	Salbuch Nordgau	
charter no.	archive ID	Monasterium.Net ID
3	SpAR Urk. 1479	12510404
4	SpAR Urk. 54	12510907
9	SpAR Urk. 1074	12530626
14	SpAR Urk. 1073	12540000
21	SpAR Urk. 395	12550803
25	SpAR Urk. 1288	12560515
26	SpAR Urk. 55a	12560617

Table 1. Sample of the tradition of the charters in urbarium *Salbuch Nordgau* listed by Kaufner (2011). Charter numbers originate from the scholarly edition.

according to places (König 2003, 23). It provides a whole set of categories[10] and places. These categories and places can be used to assign attributes to those charters which are listed in the finding aide. Additionally, the transcription of *DicMihi* lists charters that are now lost. Apart from the *DicMihi* register, the tradition of the St. Katharinenspital charters listed in various scholarly editions was also added to the data source. Initially, all documents from the printed editions were added to a temporary MySQL database with their charter number, archive IDs and Monasterium.Net IDs. Table 1 shows which of the charters in Kaufner (2011) are contained in the *Salbuch Nordgau* urbarium with their corresponding IDs.

In total, traditional references from five different scholarly editions covering 176 St. Katharinenspital charters were added to the dataset. Hence, our dataset now contains information about the tradition of charters in various historic and scholarly documents. We use this information to model the traditional context of the documents in the graph database. In conclusion, our dataset of the St. Katharinenspital is particularly suitable for the extraction of entities and relations to establish links between the charters and related documents. By using this dataset, we are already able to add a significant amount of knowledge to the rather outdated St. Katharinenspital charters collection on Monasterium.Net.

3 Extracting Entities and Relations

The heterogeneous St. Katharinenspital dataset provides the entities and relations to model cross-document interdependencies. We extract these by identifying mutual

[10] E.g., "fürstliche Privilegia", "die Spitalmühl und das Baad betreffende Brief" or "Bischöffliche Begnadigung und andere Urkunde".

entities in different individual documents, as well as by adding additional information, e.g., norm data,[11] via data normalisation efforts and natural language processing (NLP) with *spaCy*[12]. Finally, the enhanced data are modelled as an interconnected graph-structure that can be used for exploratory analyses of both the documents and the entities from many different research perspectives (e.g., historical, archival, linguistic, economic and cultural). The following types of entities are extracted or determined during the data pre-processing and analysis:

- charters (individual regests, transcripts, references…);
- legal activities (legal content of a charter…);
- actors (authors, witnesses, groups, legal entities…);
- related documents (scholarly editions, traditional documents…);
- dates (time stamps); and
- places (with or without geo-reference).

To build the data collection, the raw data are processed in an extensive data pre-processing pipeline, which consists of several individual operations. Primarily, the data are cleaned, labelled and restructured to make them machine-readable by generating *.csv, *.html or *.xlsx files. Thereby, the identified entities and relations get updated manually and automatically to be able to uniformly acquire them later on setting up the graph database. This part of the whole process is particularly time-consuming, as almost all raw data need to be sighted. Following this, normalised spellings of the entities and PIDs (archive IDs, GND references, file names etc.) are introduced. Subsequently, parts of the collected entities and relations are stored in a temporary MySQL database. Thus, the data are quickly available at different points of the preparation process, e.g., for additional metadata aggregation, such as geoparsing,[13] further data cleaning processes, and, finally, to set up the graph database (see Figure 1). Since the extraction of the entities and relations from the St. Katharinenspital dataset is experimental and the final quality of the results could not be anticipated, we decided not to store the data in an RDF-like data structure, and kept on using the MySQL database during the whole data preparation process (more on this in the following section).

We see charter abstracts as a valuable resource for the extraction of entities and relations. Therefore, we particularly focus on their analysis. Thereby, the NLP methods applied to extract entities and relations from the charter abstracts form an integral part of our data pre-processing pipeline. Charter abstracts are just a brief summary of

[11] We incorporate experimentally the Integrated Authority File (GND), managed by the German National Library (DNB), available at http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html.

[12] We use spaCy (Honnibal & Montani 2018) to extract linguistic features like part-of-speech tags, dependency labels and named entities from our corpus. Available at https://spacy.io/usage/linguistic-features.

[13] In this project, basic geo data are retrieved from GeoNames database (GeoNames 2018), in a process comparable to Jeller (2019).

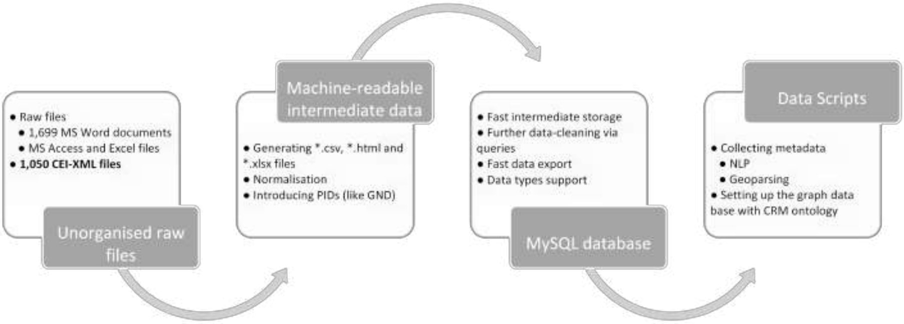

• Raw files
• 1,699 MS Word documents
• MS Access and Excel files
• 1,050 CEI-XML files
Unorganised raw files
Machine-readable intermediate data
• Generating *.csv, *.html and *.xlsx files
• Normalisation
• Introducing PIDs (like GND)
• Fast intermediate storage
• Further data-cleaning via queries
• Fast data export
• Data types support
MySQL database
Data Scripts
• Collecting metadata
• NLP
• Geoparsing
• Setting up the graph data base with CRM ontology

Figure 1. Data pre-processing pipeline. Illustration created by the authors.

the primary content of a charter and often consist of a single sentence. The following sentence is a typical example:

> "Ulrich der Frutrunch von Abbach verkauft dem St.-Katharinenspital sein Gut in Teingen (Teugn) um 15 Pfund Regensburger Pfennig."(St. Katharinenspital charter SpAR Urk. 2101)[14] *("Ulrich der Frutrunch von Abbach sells his manor in Teingen (Teugn) to the St.-Katharinenspital for 15 pounds of Regensburger Pfennig.")*

The charter abstracts are rather uniform, and thus facilitate information extraction. The approach of collecting data from German charter abstracts was already applied by Kuczera (2017): by extracting the first verb of a charter abstract and applying a lemmatizer afterwards, Kuczera creates basic lemma nodes in the graph database. The nodes are connected to the regest nodes via *HERRSCHERHANDELN* ("acts-of-rulership") edges (2017, 187). The particular focus on the verb essentially follows the theory of *dependency grammar*, according to which the dependence of a word upon another word results from a verb. This grammatical concept goes back to Lucien Tesnière (2015), and views the finite verb as the centre of the grammatical organisation of a sentence. There also exists a basic NLP script for spaCy by Georg Vogeler 2018),[15] which can be used to analyse a corpus of regests. The script automatically assigns attributes to digital regests by language determination and basic named entity extraction (NER). This script shows how relatively simple NLP techniques

[14] Regensburg, Archiv des Katharinenspitals Urkunden (1145-1568) SpAR Urk. 2101, available at https://www.monasterium.net/mom/DE-AKR/Urkunden/SpAR_Urk_2101_/charter.

[15] This is a collection of scripts and workflows for NLP experiments with data from Monasterium.Net, available at https://github.com/GVogeler/mom-NLP.

can be used to evaluate large amounts of charter regests of Monasterium.Net. More complex syntactic structures, however, are not evaluated by the script. We extend both Kuczera's and Vogeler's NLP-based approaches with a heuristic analysis of sentence structures found in German charter abstracts of the St. Katharinenspital. Essentially, we apply the spaCy default NLP pipeline (a tokenizer, a part-of-speech tagger, a dependency-parser and named entity recognition[16]) with some small modifications, such as merging identified multi-token named entities (NEs) into single tokens.[17] Analysing a charter abstract (e.g., of charter SpAR Urk. 2101[18]) with spaCy generates a *spaCy Doc object.*[19] It includes a sequence of part-of-speech tagged tokens and syntax dependencies that can be represented as follows (Figure 2):

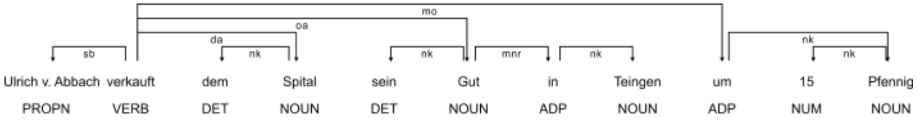


Figure 2. SpAR Urk. 2101 charter abstract syntax dependencies. Illustration created by the authors with *displaCy*[20].

In the following, a heuristic is derived from the generated dependency tree to extract the entities contained in the German language regests and their relationships to each other. The heuristic approach focuses on extracting *triples* (SVO) or *quadruples* (SVOO) of entities and verbs (nodes) as well as their mutual relations (edges) from the charter abstracts. The relations are subsequently derived from the word order of the single elements of a triple or quadruple:

```
S = subject, V = verb, O = object

SVO:['Gebolf von Ellenbach', 'verkaufen (to sale)', 'Weingut (vineyard)']
S-[:relationA]-V-[:relationB]-O

SVOO:['Rudiger der Mulnar', 'vermachen (to devise)', 'St.Katharinenspital', '
    Äcker (fields)']
S-[:relationA]-V-[:relationB]-O-[:relationC]-O
```

[16] SpaCy language processing pipelines, available at https://spacy.io/usage/processing-pipelines.

[17] We collapse consecutive tokens tagged as a named entity into a single token. Thus, a multi-token entity like "Ulrich von Abbach" that consists of tokens belonging to different lexical categories now becomes a single token with a single part-of-speech tag.

[18] SpAR Urk. 2101 on Monasterium.Net, available at https://www.monasterium.net/mom/DE-AKR/Urkunden/SpAR_Urk_2101_/charter.

[19] SpaCy Doc object definition, available at https://spacy.io/api/doc.

[20] DisplaCy Dependency Visualizer is spaCy's visualisation tool. An online demo is available at https://explosion.ai/demos/displacy.

Based on the extracted dependencies, further information may be added to the direct and indirect objects, e.g., if a direct object can be identified as a *E53Place* CRM class.[21] The heuristic is experimental and inevitably will create false positives. It heavily depends on the language model used by spaCy, as well.[22] Our heuristic basically performs four tasks for all tokens in the charter abstract. A simplified representation of the heuristic code can be seen below and is also available on GitHub:[23]

```
# heuristic to extract quadruples and triples from German charter abstracts
index = 0
subject = ''

dobject = ''
dobject2 = ''
# iterates through every token of spaCy doc object
for token in doc:
  if token.dep_ == "sb" or token.dep_ == "oa" or token.dep_ == "da":
    # task 1: get subject
    if token.dep_ == "sb" and index == 0:
      subject = token.text
    # task 2: get verb of subject (predicate)
    if token.dep_ == "sb" and index == 0 and token.head.pos_== "VERB":
      verb = token.head.lemma_
    # task 3: get direct object
    if index == 1 and (token.dep_ == "da" or token.dep_ == "oa")
                  and token.head.pos_ == "VERB":
      dobject = token.text
    # task 4: get indirect object
    if index == 2 and token.dep_ == "oa":
      dobject2 = token.text
    index += 1
if subject != '' and verb != '' and dobject != '' and dobject2 != '':
  return subject, doc, verb, dobject, dobject2
if subject != '' and verb != '' and dobject:
  return subject, doc, verb, dobject
return ''
```

The sample code shows the structural criteria that allow us to identify a structural subject, verb and object(s) in a simple German declarative phrase (Meibauer et al. 2013, 20–50). Therefore, the success of the extraction depends on the syntax of the charter abstract.[24] For the German language, we start from the *V2 word order*, which places the finite verb of a phrase or sentence in second position, with a single constituent preceding it (Haider 2010, 1f.). By using the syntax dependencies, the part-of-speech

[21] We achieve this by matching words and performing further syntactic analyses that would go beyond the scope of this publication.

[22] SpaCy's pre-trained statistical models for German, available at https://spacy.io/models/de.

[23] The implementation used is available on GitHub under GPL 3 license and can be used to generate an experimental Neo4j database with some German charter abstracts (Sippl 2019), available at https://github.com/cs-ubr/charter-abstracts.

[24] Charter abstracts in our dataset vary in length and detail. A significant part of the corpus consists of abstracts that consist of more complex syntactic structures with several phrases or even several sentences.

Dependency Label[25]	**Description**	**POS**	**Token**	**CRM Class**
sb	subject	PROPN	Ulrich v. Abbach	E21Person
-	(predicate)	VERB	verkauft (*sells*)	E7Activity
da	dative	NOUN	Spital (*St. Katharinenspital*)	E74Group
oa	accusative object	NOUN	Gut (*manor*)	E53Place

Table 2. Quadruple of subject, predicate, direct and indirect object for the charter abstract of SpAR. Urk. 2101.

tags and the word order from the spaCy Doc object, we are now able to derive the constituents of a triple or a quadruple of a phrase. For the subject and the verb this is implemented in lines 10 to 15 in the sample code. Just like Kuzcera (2017), we also use the lemmatised form of the verb. However, we use the verb to generate an entity instead of a relation (see next section). The extraction of direct and indirect objects is implemented in lines 16 to 21. If no indirect object could be extracted, only a triple consisting of the subject, the verb and the direct object is returned. Also, different spellings of actors or place names may appear in the charter abstracts. For example, *St. Katharinenspital* may occur as *St.-Katharinenspital*, *Spital*, or simply *Katharinenspital*. We therefore introduce normalised spellings for selected entities (e.g., *St. Katharinenspital*). Table 2 shows the extracted entities of the example charter that are already assigned with their corresponding CRM classes.

Table 3 shows a few example quadruples extracted from charter abstracts. The entities shown can be linked to a charter and the identified dependencies also help to describe the relationships between the entities. As can be seen, the lemmatised verbs particularly stand out, as they make it possible to categorise the charters by their legal content. The results also show that triples and quadruples are suitable to generate human readable lists. Thus, improving the data can be achieved quickly, either manually or automatically. Our heuristic identifies 225 quadruples and 407 triples in the charter data, thereby raising the number of annotated places from 292 to 1351, as well as the number of persons from 749 to 2435[26] (see Table 4 and Table 5). The results could be improved if the machine learning features of spaCy, and more detailed heuristics for different syntactic structures, were applied. The time and effort required for this, however, is too big for the collection of a small archive. In the case of much larger corpora of charter abstracts (e.g., the whole Monasterium.Net corpus),

[25] The spaCy dependency labels are based on the TIGER Treebank annotation scheme, available at https://spacy.io/api/annotation#dependency-parsing.

[26] The numbers show distinct entities. However, ambiguities and false positives remain in the data and thus, in the numbers.

Charter	grammatical subject	predicate	direct object	indirect object
SpAR_Urk_799	Rudger der Mulnar	vermachen (*to bequeath*)	St. Katharinenspital	Äcker (*fields*)
SpAR_Urk_135	Leopold von Gründlach	verleihen (*to give*)	St. Katharinenspital	Ablass (*indulgence*)
SpAR_Urk_1394	Romungus von Chamerstein	bestätigen (*to confirm*)	St. Katharinenspital	Besitz (*property*)
SpAR_Urk_1367	Otto von Dürn	verkaufen (*to sell*)	St. Katharinenspital	Hof (*farm*)
SpAR_Urk_1189	Pernger von Haydawe	übereignen (*to transfer*)	St. Katharinenspital	Wald (*forest*)
SpAR_Urk_1072a	Elspet	verkaufen (*to sell*)	St. Katharinenspital	Teil (*part*)

Table 3. Quadruples extracted from the charter abstracts by means of NLP.

Property	Value
Charters (CEI-XML files)	1,050
Persons (unique)	749
Places (unique)	292

Table 4. Entities in St. Katharinenspital Monasterium.Net dataset.

this would be a feasible approach. Additionally, our extracted data is suitable to train a *Conditional Random Fields* (CRF) classifier for improved NER results and even an automatic assignment of CRM labels. This also allows for a context sensitive extraction of numbers (e.g., quantifiers, amounts of money etc.) and thereby substantially increase the amount of extracted information. Therefore, this approach may be considered in future work. An example of a successful use of a pre-trained CRF classifier in a comparable DH context with extensive data preparation can be found in Lüschow (2020).

4 Modelling Charters and Cross-Document Interdependencies

To model the St. Katharinenspital dataset as a graph, we rely on two essential concepts: entities (nodes) and relations (edges). We use the Neo4j labelled-property graph

DB Property	Value
Nodes total	9,676
CRM-Entities (E1CRMEntity)	7,665
Edges total	24,138
Charters (E5Event)* *E7Event excluded	2,489
Persons (E21Person)	2,435
Places (E53Place)	1,351

Table 5. Entities and properties of the final labelled-property graph database.

database to store these entities and relationships. Compared to *RDF triplestores*, Neo4j supports advanced graph metrics, weighted edges and is very easy to set up and maintain from a developer's perspective. The first two aspects are relevant for network analyses, which are becoming increasingly important in DH (Jannidis et al., 2017, 147-149). The latter facilitates, as an example, the future integration of Neo4j into a large digital repository based on software for research data management such as Invenio.[27] This way, users and developers can also get around the major disadvantages of SPARQL (Vogeler 2019). In the academic world, however, RDF triplestores are widely used for storing and retrieving triples with semantic queries. A recent example for this is the project of Lüschow (2020). Therefore, the migration of our data into an RDF database, after an extensive data evaluation process, is a future application. However, a detailed comparison of RDF and Neo4j would go beyond the scope of this paper, especially since it is also the subject of an ongoing debate[28].

The extracted entities can take various forms, including persons, institutions or places. Kuczera's (2017) example of verb extraction and the modelling of relationships between regests and persons as *HERRSCHERHANDELN* ("acts-of-rulership") edges in the graph is ultimately a description of a historic event. According to our interpretation, the extracted verb describes this abstract event, which in turn can be assigned to a date, a person, an object that is changed by the event and a specific document (Figure 3).

Events are a central concept in modelling data from domains such as history or cultural heritage (Van Hage et al. 2011). However, the granularity of a data model,

[27] Invenio is an Open Source framework for large-scale digital repositories. It was initially developed by CERN and features support for large-scale research data managmenet use cases, available at https://invenio-software.org/.

[28] E.g., this article on the technology news website ZDNet contrasts contrary statements of Neo4j's CEO with the statements of an advanced GraphDB user, thus highlighting the technological and ideological differences between the two approaches, available at https://www.zdnet.com/article/graph-databases-and-rdf-its-a-family-affair/.

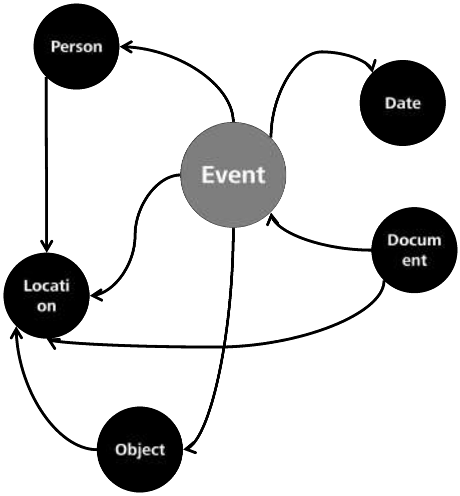


Figure 3. Event as central element of a graph-based data model. Illustration created by the authors.

the data to be captured, and the essential components of it, ultimately depend on the overall requirements a data model needs to meet. As we describe relations between entities by incorporating and linking information from CEI-XML files, Microsoft Office files, archival metadata, biographical data, place attributions, roles, basic prosopographic knowledge or data we generate with NLP, we require a data model that allows us to map basic forms of traditional diplomatic work. This data model needs to be sustainably implemented in a versatile technical environment, as users (institutions, students, scholars and software developers) have different demands. CIDOC-CRM is a model that can be used within graph databases, and it is particularly suitable if complex problems can't be solved in a monodisciplinary manner. It is an ISO standard and intended to enable discipline-independent linking of heterogeneous cultural information, especially from cultural heritage stored in museums, libraries or archives. For this purpose, logically defined terms – classes and properties – are used. Both are intended to support a great number of information resources (Le Bœuf et al. 2015, ii). Apart from using the ontology for medieval charters (Ore 2009), there have been proposals to use CIDOC-CRM for account books as well (Vogeler 2015), thus further primary sources from the St. Katharinenspital archive (e.g., account books or urbaria) can easily be added to the model in future applications. A factoid-based ontology expressed by the CIDOC-CRM model was also proposed to improve interoperability, and enable better support of prosopographic databases (Pasin & Bradley 2015). This outlines the fact that very different source types can be incorporated in a single data

Nodes (CRM classes)	Edges (CRM properties) connected to E5Event
E7Activity	P20HadSpecificPurpose P9ConsistsOf
E21Person	P11HadParticipant
E30Right	P129IsAbout
E31Document	P70Documents
E52TimeSpan	P4HasTimeSpan
E53Place	P7TookPlaceAt P161HasSpatialProjection
E55Type	P2HasType
E74Group	P11HadParticipant

Table 6. Edges of *E5Event* in the data model.

model, and how well it is suited to the project. Central classes of the ontology are *E5Event*[29], thing *E70Thing*[30] and *E39Actor*[31]. To better illustrate how the CRM classes can be extracted from charter abstracts, they are annotated in the following for charter SpAR Urk. 2101:

> [E21Person: *Ulrich der Frutrunch von Abbach*] [E7Activity: *verkauft (sells)*] dem [E74Group: *St.-Katharinenspital*] sein [E53Place: *Gut (manor)*] in [E53Place: *Teingen (Teugn)*] um 15 Pfund Regensburger Pfennig. (St. Katharinenspital charter SpAR Urk. 2101)

As described in the previous section, entities and relations are derived from the syntactic structure of the regests abstracts using our spaCy NLP heuristic. The resulting data model is based on the CIDOC-CRM v6.2.1. Thereby, the following CRM relationships are used to connect a charter (*E5Event*) to other CRM entities (Table 6):

The nodes (*E1CRMEntities*) in our data model may be connected via different edges to a charter (*E5Event*). The easiest way to understand this is by looking at the extracted quadruples. A charter as an instance of CRM class *E5Event* is at the centre of our data mode. As a result, the quadruple for the example charter consists of `"SpAR Urk. 2101"` (*E5Event*), `"Ulrich v. Abbach"` (*E21Person*), `"verkaufen"` *E7Activity*), `"Spital"` (*E39Group*), and `"Gut"` (*E53Place*) class instances, and is represented within the graph database as illustrated in Figure 4.

[29] "This class comprises changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological or legal phenomena." (Le Bœuf et al. 2015, 5).

[30] "This general class comprises discrete, identifiable, instances of *E77PersistentItem* that are documented as single units, that either consist of matter or depend on being carried by matter and are characterized by relative stability." (Le Bœuf et al. 2015, 70).

[31] "This class comprises people, either individually or in groups, who have the potential to perform intentional actions of kinds for which someone may be held responsible." (Le Bœuf et al. 2015, 20).

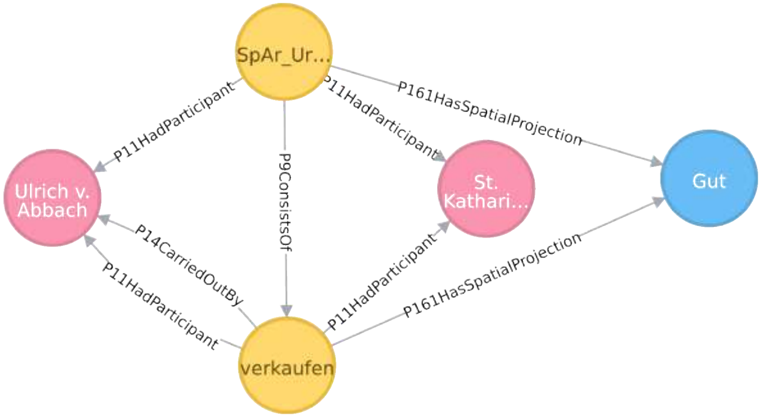


Figure 4. Graph representation of charter abstract SpAR Urk. 2101. Illustration created by the authors with Neo4j Browser.

The visualisation tool of the Neo4j data base allows for the easy creation of graph views, thus the overall properties of the generated data model can be quickly checked. The nodes (entities) are connected via various edges that represent different semantic relationships. Since the example given is a *property sale*, there is a *seller* (Ulrich v. Abbach) and a *buyer* (St. Katharinenspital). They participate in the transaction *E7Activity*, so both seller and buyer are connected via a *P11HadParticipant*[32] CRM property to the *E7Activity*. As the charter is represented as an instance of *E5Event*, both actors are also connected to it via a *P11HadParticipant* edge. As an abstract event, the charter describes the property sale. Hence, it consists of (*P9ConsistsOf*[33]) the property sale. This logic can be applied to all relationships depicted in Figure 4. If all nodes are connected correctly, five nodes with eight edges are created in the database from just a single quadruple. The CIDOC-CRM constraints and structure are enforced with the *cidoc-crm-neo4j* script[34], so that every CRM subclass has the labels of its super classes, when created in the database. Also, the validity of the data model is guaranteed by enforcing the right relations between individual nodes and inhibiting relations that are contrary to the CIDOC-CRM definition. Hence, the

[32] "This property describes the active or passive participation of instances of E39 Actors in an E5 Event." (Le Bœuf et al. 2015, 48).

[33] "This property associates an instance of E4 Period with another instance of E4 Period that is defined by a subset of the phenomena that define the former." (Le Bœuf et al. 2015, 47).

[34] The CIDOC-CRM constraints and class hierarchies are enforced with *cidoc-crm-neo4j*, a python-based script by Erick Peirson, ASU, GPL3 (2014), which in turn relies on the latest RDFS serialisation of CIDOC-CRM (v6.2.1), available at http://www.cidoc-crm.org/sites/default/files/cidoc_crm_v6.2.1-2018April.rdfs.

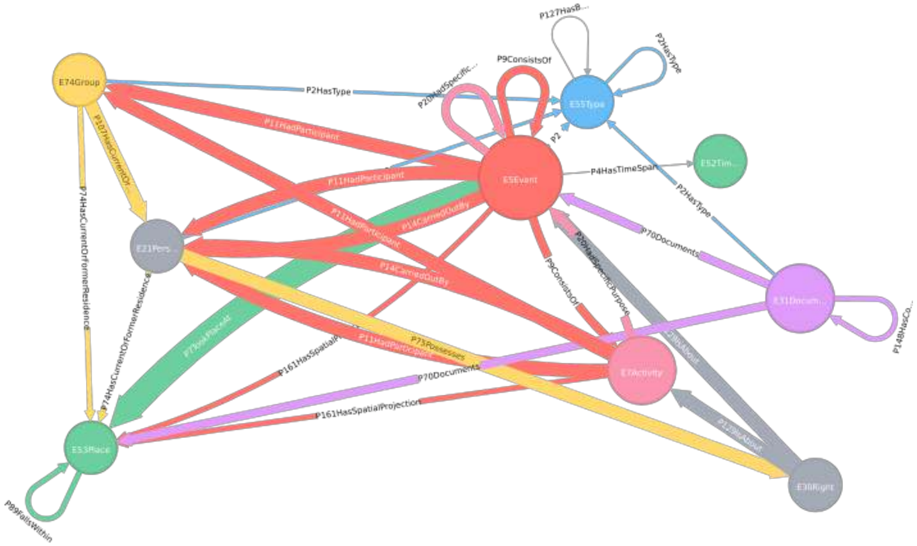


Figure 5. CIDOC-CRM data model representation. Illustration created by the authors with Neo4j Browser.

simplified data model as described in Table 6 can be illustrated as a complex graph structure (see Figure 5).

With the help of the data model implementation in the Neo4j database, custom *cypher queries*[35] can now be executed for all CRM subclasses, super classes and properties. Which charters in the dataset are type of *Bischofsurkunde* (*episcopal deeds*)? The following cypher queries produce the same results:

```
MATCH (n:E1CrmEntity)-[]-(m:E1CrmEntity) WHERE m.name="Bischofsurkunde"
RETURN n.spa_id

MATCH (n:E5Event)-[:P2HasType]-(m:E55Type) WHERE m.name="Bischofsurkunde"
RETURN n.spa_id
```

Both queries return 109 archive ids. Since all nodes are instances of the class *E1CRMEntity*[36] and only charters may have a connection to *E55Type*[37] *Bischofsurkunde*, the query results are identical. In the digital regests of the St. Katharinenspital on Monasterium.Net, up until now it has not been possible to specifically search for

[35] Cypher is Neo4j's graph query language, further information in Neo4j (2018).

[36] "This class comprises all things in the universe of discourse of the CIDOC Conceptual Reference Model." (Le Bœuf et al. 2015, 2).

[37] "This class comprises concepts denoted by terms from thesauri and controlled vocabularies used to characterize and classify instances of CRM classes." (Le Bœuf et al. 2015, 27).

actors mentioned in a charter – if they had been annotated there – or the kind of relationship they have to a charter document.

```
MATCH (n:E5Event)-[]-(m:E21Person) WHERE m.name="Otto Prager" RETURN n.spa_id
"SpAr_Urk_86"
"SpAr_Urk_35"
"SpAr_Urk_85"
"SpAr_Urk_691"
"SpAr_Urk_692"
"SpAr_Urk_1128"
"SpAr_Urk_469"
"SpAr_Urk_76"
"SpAr_Urk_54"
"SpAr_Urk_466"
"SpAr_Urk_75"
```

The query result shows all charters in the database where *Otto Prager* is annotated as an actor. This query can now be extended, for example by an *E7Activity*, i.e., the legal form: What is the legal form of the charters in which Otto Prager is involved?

```
MATCH (a:E7Activity)-[]-(n:E5Event)-[]-(m:E21Person) WHERE m.name="Otto Prager"
    RETURN DISTINCT n.spa_id, a.name
"SpAr_Urk_86""verkaufen" (to sell)
"SpAr_Urk_85""bestätigen" (to confirm)
"SpAr_Urk_691""bestätigen" (to confirm)
"SpAr_Urk_692""verkaufen" (to sell)
"SpAr_Urk_1128""bestätigen" (to confirm)
"SpAr_Urk_1128""verpflichten" (to oblige)
"SpAr_Urk_469""schenken" (to gift)
"SpAr_Urk_76""schenken" (to gift)
"SpAr_Urk_54""verkaufen" (to sell)
"SpAr_Urk_75""überlassen" (to convey)
```

The results show that data generated by rather simple cypher queries may already provide new insights. In addition, the graph makes it possible to reveal indirect relationships between nodes. In the context of social media, this is often referred to as *Friend of a Friend* (FOAF). In our data model, the actors (*E39Actor*) are linked indirectly as witnesses or judicial actors via charters (*E5Events*) and their legal content (*E7Activity*). This interconnection represents cross-document dependencies of the St. Katharinenspital charters and is therefore particularly interesting for a closer examination. With the following query a subgraph is generated, which can be examined more closely. The Cypher query is as follows:

```
MATCH p=(n:E39Actor)-[]-(c:E5Event)-[]-(m:E39Actor)
     WHERE NOT c:E7Activity RETURN p
```

The query result consists of 662 nodes connected via 626 edges. For this quantity of nodes, a visualisation is useful. Figure 6 shows a part of the resulting subgraph. It is particularly interesting to see that there are numerous nodes that are not linked to one another outside of a very dense network. This highlights the fact that our data basis is indeed very irregular, due to the heterogeneous states of primary source analysis, and since some charter abstracts only bare very little additional information

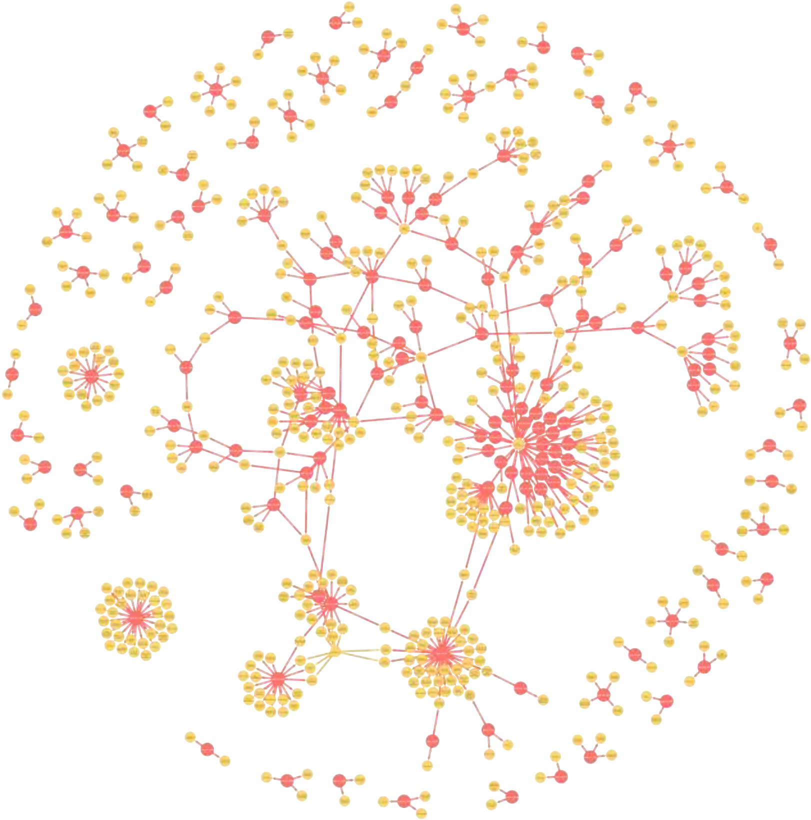


Figure 6. Linking of *E39Actors* via *E5Event* in the graph database. The network shows clearly the separation between charters not linked to any other and a very dense cluster. Illustration created by the authors with Neo4j Browser.

that can be extracted. Nevertheless, this result shows the extent to which the graph database can be used in an application that represents the various contexts of an entity, regardless of whether it is a charter, an actor or a place.

5 Creating a Web-Based Charter Portal

Charter platforms like Monasterium.Net are typically focussed on retrieving and analysing single documents. By contrast, our data model allows for further exploratory analyses (Warwick 2012, 2). We provide a web application that combines the capacities of a full-text search[38] with queries to our graph database. Some of these queries have been discussed in the previous section. The aim of the web portal accessible at https://urkunden.ur.de is to demonstrate some the capabilities of our ontology-based data model in an application that can be used by anybody. Entities that are associated with the charters – like actors (*E39Actor*), places (*E53Place*) or documents (*E31Document*) – are shown to the user and can be accessed via persistent hyperlinks. Since the charter portal is a technical demonstration and the database is not validated yet, it should not be used like a research platform. It could, however, be used as a starting point into research related to the St. Katharinenspital charters, as it delivers a quick overview of all the data that are related to the charters. Each entity (CRM class instance) is displayed in its unique context, i.e., it is shown together with its neighbours, exactly as in the graph. We use this to enhance the result lists of a full-text search (see the coloured badges in Figure 7) and to generate entity-specific detail pages (Figure 8). The detail pages, e.g., for charters like the one shown in Figure 8,[39] show the context of the charter, which consists of related actors (*E21Person*), traditional documents, as well as other charters, that are connected indirectly over common edges to *E21Person* entities (i.e., friends of a friend). Furthermore, a hyperlink to the corresponding CEI-XML file on Monasterium.Net gives users a reference to a source that can be cited. The data model based on CIDOC-CRM also allows for the creation of class overview pages, where users get a view of all instances of a CRM class and, at the same time, traverse the CRM class hierarchy (see the yellow badge in Figure 9). Thereby, registers or entire lists of entities are created automatically via queries to the graph database.[40]

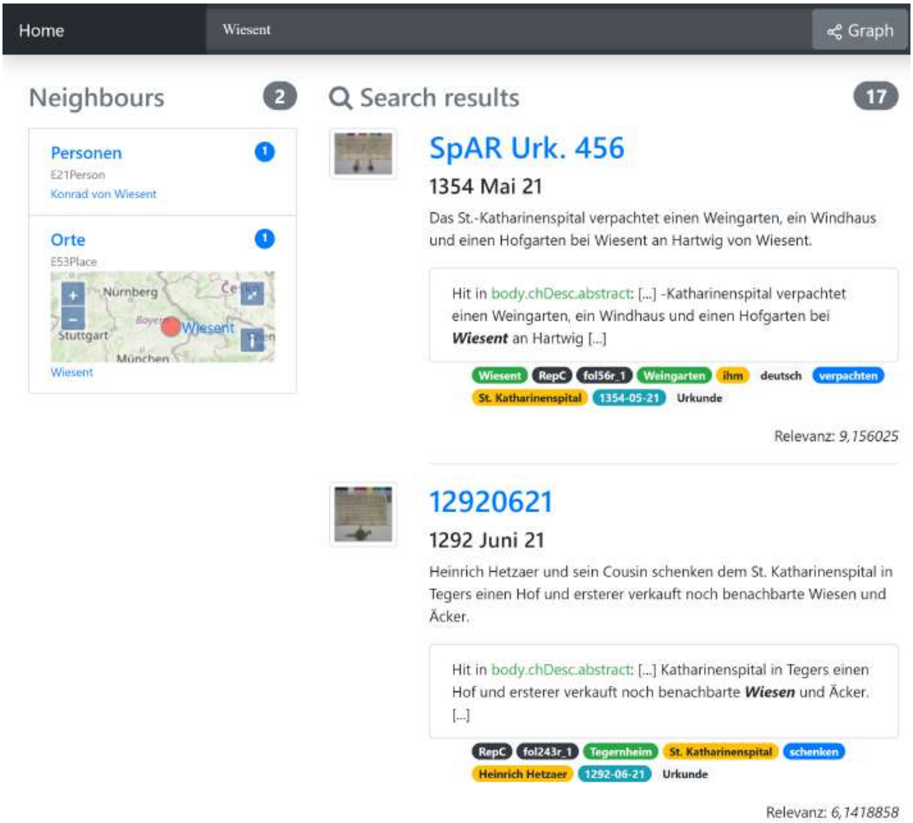
6 Conclusion

Our graph-based data model facilitates the analysis of higher-level relationships between entities and documents as it supports complex and far-reaching data queries. In our web application, users can conveniently use these queries to exploratively search the graph-based St. Katharinenspital data collection. As the graph database is traversed it delivers context-based results. The context of an entity node like a charter or an actor arises from its connection to other nodes and through its classification with the CIDOC-CRM ontology. Hence, our graph-based digital edition opens entirely new research perspectives, as the various attributes of a charter, e.g., the places of its creation, related witnesses, or traditional documents, are immediately available to

[38] For full text indexing we use Elasticsearch (Elastic 2019).

[39] CRM class *E5Event* detail page for SpAR Urk. 54 on the charter portal, available at https://urkunden.ur.de/index.php?crmentity=E5Event&prop=name&val=SpAr_Urk_54.

[40] E.g. this *E52TimeSpan* overview page lists all dated charters by their year of creation, available at https://urkunden.ur.de/index.php?crmentity=E52TimeSpan.



SpAR Urk. 456

1354 Mai 21

Das St.-Katharinenspital verpachtet einen Weingarten, ein Windhaus und einen Hofgarten bei Wiesent an Hartwig von Wiesent.

Hit in body.chDesc.abstract: [...] -Katharinenspital verpachtet einen Weingarten, ein Windhaus und einen Hofgarten bei ***Wiesent*** an Hartwig [...]

Relevanz: *9,156025*

12920621

1292 Juni 21

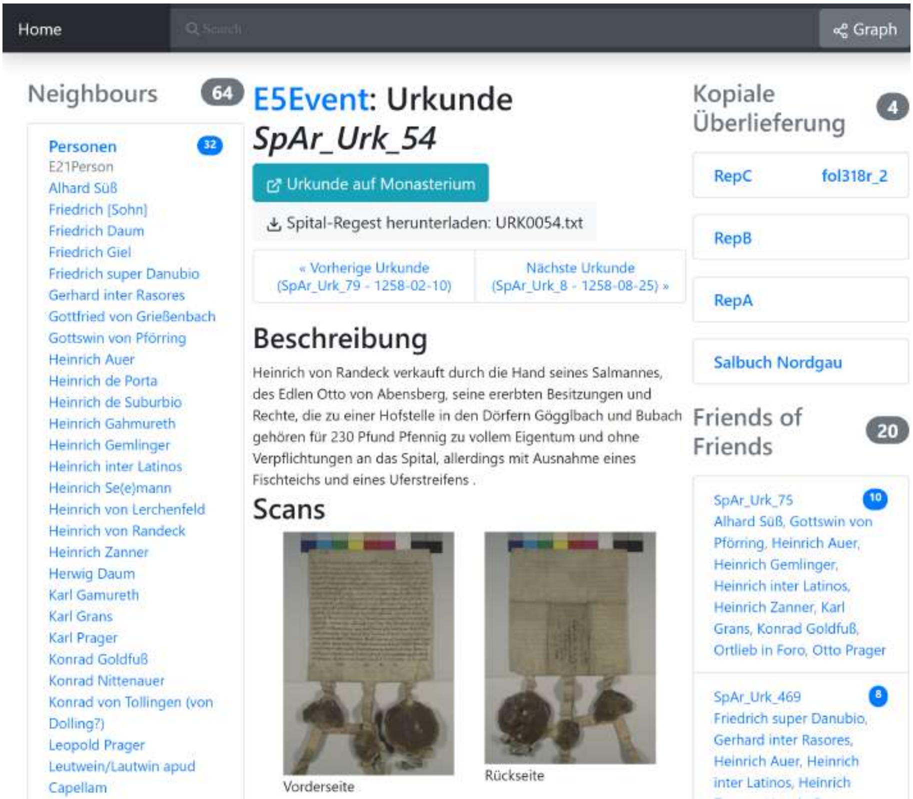
Heinrich Hetzaer und sein Cousin schenken dem St. Katharinenspital in Tegers einen Hof und ersterer verkauft noch benachbarte Wiesen und Äcker.

Hit in body.chDesc.abstract: [...] Katharinenspital in Tegers einen Hof und ersterer verkauft noch benachbarte ***Wiesen*** und Äcker. [...]

Relevanz: *6,1418858*

Figure 7. Full text search results for *Wiesent* (place in Bavaria). Screenshot created by the authors.

the user, jointly with their CIDOC-CRM labels. These different entities provide the facets for an experimental search portal with fulltext search capabilities. Thus, users may quickly get an overview of an entire collection of charters, and receive more information, compared to a basic fulltext search. By overcoming the limits of CEI-XML mark-up in the graph, we have created a data structure that, furthermore, facilitates the incorporation of external data sources like authority files or geospatial information. Besides this, our database can now be easily imported by charter platforms like Monasterium.Net and be used to help enhance their data basis. Further on, we showed how the application of NLP procedures facilitates the extraction of large quantities of entities from corpora of charter abstracts. We achieved this by developing a simple



Neighbours 64

Personen 32
E21Person
Alhard Süß
Friedrich [Sohn]
Friedrich Daum
Friedrich Giel
Friedrich super Danubio
Gerhard inter Rasores
Gottfried von Grießenbach
Gottswin von Pförring
Heinrich Auer
Heinrich de Porta
Heinrich de Suburbio
Heinrich Gahmureth
Heinrich Gemlinger
Heinrich inter Latinos
Heinrich Se(e)mann
Heinrich von Lerchenfeld
Heinrich von Randeck
Heinrich Zanner
Herwig Daum
Karl Gamureth
Karl Grans
Karl Prager
Konrad Goldfuß
Konrad Nittenauer
Konrad von Tollingen (von Dolling?)
Leopold Prager
Leutwein/Lautwin apud Capellam

E5Event: Urkunde *SpAr_Urk_54*

Urkunde auf Monasterium

Spital-Regest herunterladen: URK0054.txt

« Vorherige Urkunde (SpAr_Urk_79 - 1258-02-10)

Nächste Urkunde (SpAr_Urk_8 - 1258-08-25) »

Beschreibung

Heinrich von Randeck verkauft durch die Hand seines Salmannes, des Edlen Otto von Abensberg, seine ererbten Besitzungen und Rechte, die zu einer Hofstelle in den Dörfern Gögglbach und Bubach gehören für 230 Pfund Pfennig zu vollem Eigentum und ohne Verpflichtungen an das Spital, allerdings mit Ausnahme eines Fischteichs und eines Uferstreifens .

Scans

Vorderseite

Rückseite

Kopiale Überlieferung 4

RepC fol318r_2

RepB

RepA

Salbuch Nordgau

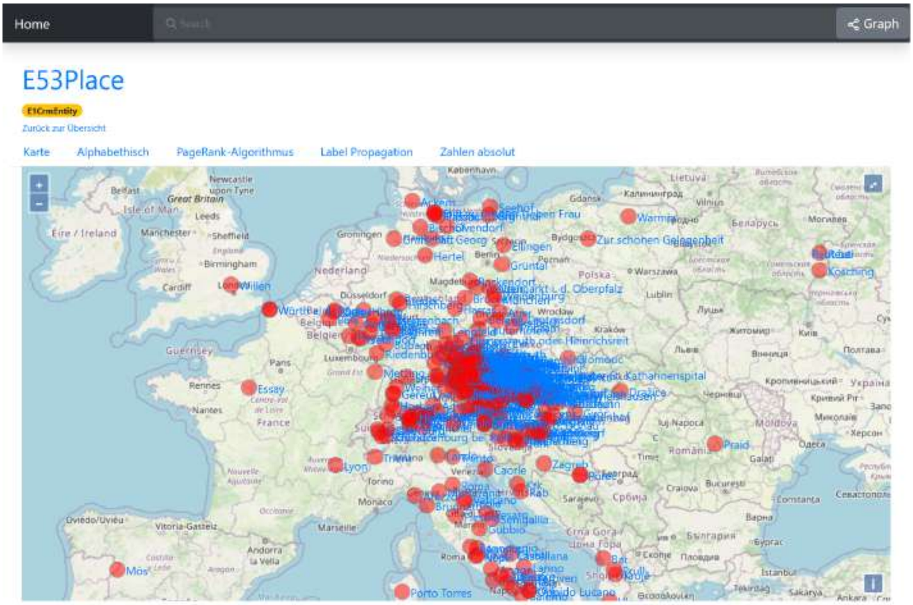
Friends of Friends 20

SpAr_Urk_75 10
Alhard Süß, Gottswin von Pförring, Heinrich Auer, Heinrich Gemlinger, Heinrich inter Latinos, Heinrich Zanner, Karl Grans, Konrad Goldfuß, Ortlieb in Foro, Otto Prager

SpAr_Urk_469 8
Friedrich super Danubio, Gerhard inter Rasores, Heinrich Auer, Heinrich inter Latinos, Heinrich

Figure 8. Graph based charter SpAR Urk. 54 and its next neighbours (CRM class *E5Event* detail page). Screenshot created by the authors.

heuristic approach. However, spaCy provides machine learning features that could provide better results, and so help to analyse more complex syntactic structures. Beyond this, our extracted data can be used to train a CRF model for improved NER in future applications. This is particularly interesting for the analysis of entire data collections e.g., on Monasterium.Net or other charter platforms. Finally, the examination of higher-level correlations by means of quantitative network analysis methods is an established approach. Therefore, a deeper analysis of graph structures is an integral part of future scenarios that could be applied to a broader dataset. Thus, the structure and the content of our graph database is to be viewed as a not-yet-completed, open-ended process.



E53Place

E1CrmEntity

Zurück zur Übersicht

Karte Alphabethisch PageRank-Algorithmus Label Propagation Zahlen absolut

Figure 9. CRM class *E53Place* overview page[41]. Screenshot created by the authors.

Acknowledgements

We would like to thank the archive of the St. Katharinenspital in Regensburg for making the data sources available to us. Special thanks to Dr. Gernot Deinzer from the University Library of Regensburg, who provided the server infrastructure and supported our project. We would also like to thank both Dr. Žarko Vujošević from the Faculty of Philosophy of the University of Belgrade and ICARUS, who made it possible for us to present the project to an international audience at short notice.

Bibliography

Ambronn, Karl-Otto, *Verwaltung, Kanzlei und Urkundenwesen der Reichsstadt Regensburg im 13. Jahrhundert*, Münchener historische Studien (Kallmünz, Opf.: Lassleben, 1968)

Deutsche Nationalbibliothek, *Gemeinsame Normdatei (GND)*, 2018 <https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html>

[41] CRM class *E53Place* overview page. Available at https://urkunden.ur.de/index.php?crmentity=E53Place.

Elastic, *Elasticsearch*, version 6.7.2 (Elasticsearch B.V., 2019) <https://www.elastic.co/de/products/elasticsearch>

Feichtmeier, Simon, *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1296 - 1301)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals et al., [preprint])

FORTH-ICS, *CIDOC CRM v6.2.1 (Draft) Encoded in RDFS* (Athen: FORTH-ICS, 2018) <http://www.cidoc-crm.org/sites/default/files/cidoc_crm_v6.2.1-2018April.rdfs>

GeoNames, *GeoNames Gazetteer*, 2018 <http://download.geonames.org/export/>

Haider, Hubert, *The Syntax of German*, Cambridge Syntax Guides, 1. publ. (Cambridge: Cambridge Univ. Press, 2010)

Honnibal, Matthew, and Ines Montani, *SpaCy - Industrial-Strength Natural Language Processing in Python*, version 2.0.16 (Berlin: ExplosionAI GmbH, 2018) <https://spacy.io/>

Jannidis, Fotis, Hubertus Kohle, and Malte Rehbein, eds., *Digital Humanities: eine Einführung* (Stuttgart: J.B. Metzler Verlag, 2017)

Jeller, Daniel, 'Urkunden als Netzwerk. Ein Werkstattbericht.', in *Quellen, Nachbarschaft, Gemeinschaft. Auf dem Weg zu einer gemeinsamen Kulturgeschichte Zentraleuropas*, ed. by Adelheid Krah (Wien; Köln; Weimar: Böhlau Verlag, 2019), 84–95 <https://dighist.hypotheses.org/945>

Kaufner, Dominik A., *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1251 - 1258)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals Regensburg, 2011)

König, Stefan, *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1145 - 1251)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals Regensburg, 2003)

Kuczera, Andreas, 'Digital Editions beyond XML – Graph-Based Digital Editions', in *HistoInformatics 2016 - The 3rd HistoInformatics Workshop. Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016)*, 2016, 37–46 <http://ceur-ws.org/Vol-1632/paper_5.pdf>

———, 'Graphentechnologien in den Digitalen Geisteswissenschaften', *ABI Technik*, 37.3 (2017), 179–196 <https://doi.org/10.1515/abitech-2017-0042>

Le Bœuf, Patrick, Martin Doerr, Christian Emil Ore, and Stephen Stead, *Definition of the CIDOC Conceptual Reference Model. Version 6.2.1.*, 2015 <http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.1.pdf>

Lüschow, Andreas, 'Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane', in *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, ed. by Christof Schöch (presented at the DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation, Parderborn, 2020), 80–84 <https://doi.org/10.5281/zenodo.3666690>

Meibauer, Jörg, Markus Steinbach, and Hans Altmann, eds., *Satztypen des Deutschen*, De Gruyter Lexikon (Berlin ; Boston: De Gruyter, 2013)

Neo4j, *Cypher Query Language Reference*, version 9, [2018] <https://s3.amazonaws.com/artifacts.opencypher.org/openCypher9.pdf>

———, *Neo4j Community Edition*, version 3.4.0 (Neo4j, Inc., 2018) <https://neo4j.com/>

Ore, Christian Emil, 'New Digital Assets - How to Integrate Them?', in *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*, Archiv Für Diplomatik, Schriftgeschichte, Siegel- Und Wappenkunde (Köln [u.a.]: Böhlau, 2009), xii, 238 – 254

Pasin, Michele, and John Bradley, 'Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach', *Literary and Linguistic Computing*, 30.1 (2015), 86–97 <https://doi.org/10.1093/llc/fqt037>

Peirson, Erick, *Cidoc-Crm-Neo4j*, version 0.1, 2017 <https://github.com/diging/cidoc-crm-neo4j>

Sahle, Patrick, *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung*, Schriften des Instituts für Dokumentologie und Editorik, 3 vols (Norderstedt: Books on Demand, 2013)

Sippl, Colin, *Charter-Abstracts*, 2019 <https://github.com/cs-ubr/charter-abstracts>

Sturm, Ferdinand, *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1259 - 1270)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals, 2013)

Tesnière, Lucien, *Elements of Structural Syntax* (Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2015)

Van Hage, W. R., V. Malaisé, R. Segers, L. Hollink, and G. Schreiber, 'Design and Use of the Simple Event Model', *Web Semantics: Science, Services and Agents on the World Wide Web*, 9.2 (2011), 128–136

Vogeler, Georg (ed.), *CEI - Charters Encoding Initiative* (München, 2004) <http://www.cei.lmu.de/>

———, *Mom-NLP*, 2018 <https://github.com/GVogeler/mom-NLP>

———, 'Von IIIF Zu IPIF? Ein Vorschlag für den Datenaustausch über Personen', in *DHd 2019 Digital Humanities: Multimedial & Multimodal. Konferenzabstracts*, ed. by Patrick Sahle (presented at the DHd 2019 Digital Humanities: multimedial & multimodal, Frankfurt am Main, 2019), 238–41 <https://doi.org/10.5281/zenodo.2600812>

———, 'Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?', *Grenzen Und Möglichkeiten Der Digital Humanities*, 1. Sonderband der Zeitschrift für digitale Geisteswissenschaften (2015), text/html Format <https://doi.org/10.17175/sb001_007>

Warwick, Claire, 'Studying Users in Digital Humanities', *Digital Humanities in Practice*, 17.5 (2012), 1–21 <https://doi.org/10.1.1.305.6694>

Appendices

Biographical Notes

Thomas Ahrend (University of Basel, Switzerland – thomas.ahrend@unibas.ch) studied Musicology, Philosophy and Literary Studies in Frankfurt a. M. and Berlin. He received his MA 1996, and his PhD 2005 at Technische Universität Berlin with a dissertation on the instrumental music of Hanns Eisler. 1997–2010 member of the editorial staff of the Hanns Eisler Gesamtausgabe in Berlin. Since September 2010, member of the editorial staff of the Anton Webern Gesamtausgabe at Musikwissenschaftliches Seminar at University of Basel.

Peter Boot (Huygens ING, The Netherlands – peter.boot@huygens.knaw.nl) studied mathematics and Dutch language and literature; he wrote his PhD thesis about annotation in scholarly digital editions and its implications for humanities scholarship. He oversaw the creation of the digital edition of the letters of Vincent van Gogh. He is employed as a senior researcher at the Huygens Institute for the History of the Netherlands where he works, among other things, as a consultant in several edition projects.

Manuel Burghardt (University of Leipzig, Germany – burghardt@informatik.uni-leipzig.de) is head of the Computational Humanities Group at Leipzig University. He is interested in the use of digital tools and computational techniques to explore new modes of doing research in the humanities. His most recent areas of research are Sentiment Analysis in the Humanities, Drametrics, Computational Intertextuality, Computational Analysis of Movies and Series and Music Information Retrieval.

Toby Burrows (University of Oxford, United Kingdom – toby.burrows@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre at the University of Oxford, and a Senior Honorary Research Fellow in the School of Humanities at the University of Western Australia.

Hugh Cayless (Duke University, USA - hugh.cayless@duke.edu) is Senior Digital Humanities Developer at the Duke Collaboratory for Classics Computing. Hugh has over a decade of software engineering expertise in both academic and industrial settings. He also holds a Ph.D. in Classics and a Master's in Information Science. He is one of the founders of the EpiDoc collaborative and currently serves on the Technical Council of the Text Encoding Initiative.

Hans Cools (University of Basel, Switzerland – 1961-2021) had a master degree in medicine and a specialization in orthopaedic surgery and traumatology (Universities of Ghent and Antwerp, Belgium, 1997), a bachelor's degree in physical therapy, and a standalone degree in informatics (1999). Through various research and project management positions, in both companies and academic institutions, he gained expertise in different aspects of the Semantic Web technologies, focusing particularly on formal data modeling and machine reasoning. Those positions were in internationally collaborative research projects in a biomedical setting, mainly of the 5-7th EU Framework Program. Foremost in these projects were semantic interoperability and reusability of data. Since 2016, he worked in the humanities, as knowledge engineer, ontologist, and Semantic Web technology expert, at the University of Basel, as part of the NIE-INE project, which highlights scholarly editing. He (co-)published several articles, and gave workshops on the implementation of Semantic Web technologies in biomedicine and the humanities. He passed away in April 2021.

Francesca Giovannetti (University of Bologna, Italy – francesc.giovannett6@unibo.it) is a second-year PhD student in Digital Humanities at the Department of Classical Philology and Italian Studies, University of Bologna. She received an MA in Digital Humanities from King's College London and a second cycle degree in Digital Humanities and Digital Knowledge from the University of Bologna. She is interested in combining digital scholarly editing with semantic web technologies and in the use of digital technologies in education.

Matthew Holford (University of Oxford, United Kingdom – matthew.holford@bodleian.ox.ac.uk) is Tolkien Curator of Medieval Manuscripts at the Bodleian Library, University of Oxford.

Marijn Koolen (Royal Netherlands Academy of Arts and Sciences - Humanities Cluster, The Netherlands – marijn.koolen@gmail.coml) studied artificial intelligence and wrote his PhD thesis on using hyperlinks in information retrieval algorithms. He has worked on scholarly annotation for digital humanities research and on annotation-related information behaviour and information systems. He works as a researcher and developer at the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences, where he leads a project on developing annotation support within the *CLARIAH research infrastructure* project.

David Lewis (University of Oxford, United Kingdom – david.lewis@oerc.ox.ac.uk) is a Research Associate in the Oxford e-Research Centre at the University of Oxford.

Andrew Morrison (University of Oxford, United Kingdom – andrew.morrison@bodleian.ox.ac.uk) is a Software Engineer in the Bodleian Digital Library Systems and Services, Bodleian Library, University of Oxford.

Stefan Münnich (University of Basel, Switzerland – stefan.muennich@unibas.ch) studied musicology and communication science at the Technische Universität Berlin, MA 2011 with a thesis on cantional setting in Heinrich Schütz's Becker-Psalter. 2012 research assistant, 2013–2015 research associate of the Felix Mendelssohn Bartholdy. Sämtliche Briefe edition at University of Leipzig (co-editor of vols. 9 & 12). Since October 2015 research associate of the Anton Webern Gesamtausgabe, Basel; received his Doctorate degree in 2020 at the department of musicology at the University of Basel with a dissertation about music notation and its codes.

Iian Neill (Digital Academy of the Academy of Sciences and Literature, University of Mainz - Iian.Neill@adwmainz.de) is a visiting researcher at the Digital Academy of the Academy of Sciences and Literature Department at the University of Mainz, Germany. He is the creator of Codex, a text annotation environment which uses standoff property annotation to generate entities in a graph meta-model. Codex is currently being used to produce a digital edition of the epistles of Hildegard von Bingen at the Digital Academy in Mainz.

Roberta Padlina (University of Basel, Switzerland – roberta.padlina@unibas.ch) studied medieval philosophy at the University of Fribourg, Switzerland, obtaining a doctoral degree in June 2020. She has twelve years of professional experience in the field of Digital Humanities, thanks to which she has been able to work closely with different actors involved in the online publication of open access research. Roberta has worked for several years for e-codices –Virtual Library of Manuscripts in Switzerland and currently coordinates the National Infrastructure for Editions (NIE-INE) project. Roberta's main focus is on the opportunities and challenges that the digital shift poses for traditional education and research institutions, including developing semantic web strategies for scholarly publications and cultural goods.

Kevin Page (University of Oxford, United Kingdom – kevin.page@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre and Associate Member of Faculty in the Department of Engineering in the University of Oxford.

Miller C. Prosser (University of Chicago, USA – m-prosser@uchicago.edu) earned his Ph.D. in Northwest Semitic Philology from the University of Chicago. His academic interests include the social and economic structure of Late Bronze Age Ras Shamra-Ugarit and the use of computational methods for philological and archaeological research. Miller is the Associate Director of the Digital Studies MA program at the University of Chicago where he teaches courses on Data Management and Data Publication for the Humanities. He also works as a researcher at the OCHRE Data Service of the Oriental Institute of the University of Chicago where he consults with and supports research projects using the Online Cultural and Historical Research Environment (OCHRE). He has also worked as a tablet photographer for the Mission de Ras Shamra (Ugarit) and the Persepolis Fortification Archive Project, employing advanced digital photographic methods such as reflectance transformation imaging, photogrammetry, and high-resolution digital scanning.

Matteo Romanello (Université de Lausanne, Switzerland - matteo.romanello@unil.ch) is Ambizione SNF Lecturer at the University of Lausanne, where he conducts a project on the commentary tradition of Sophocles' Ajax. Matteo is a Classicist and a Digital Humanities specialist with expertise in various areas of the Humanities, including archaeology and history. After obtaining his PhD from King's College London, he worked as a research scientist at EPFL's DHLAB on the Linked Books and Impresso projects, before moving to his current position. He was also teaching fellow at the University of Rostock, researcher at the German Archaeological Institute, and visiting research scholar at Tufts University.

Sandra Schloen (University of Chicago, USA – sschloen@uchicago.edu) is the Manager of the OCHRE Data Service at the Oriental Institute of the University of Chicago, and is the co-designer and developer of the Online Cultural and Historical Research Environment (OCHRE). Trained in computer science and mathematics (B.Sc. University of Toronto; M.Ed. Harvard University), Sandra has spent over 30 years working with technology as a systems analyst, technical trainer, and software developer. A long association with colleagues in the academic community has enabled her to develop a specialty in solving problems in the Digital Humanities where challenges of data capture, data representation and data management abound. Specifically, she has served extensively as a database manager for several archaeological projects in Israel and Turkey, and supports a wide range of research projects at the Oriental Institute and at other universities.

Desmond Schmidt (University of Bologna - desmond.allan.schmidt@gmail.com) has a background in classical Greek philology, information security and eResearch. He has worked on several scholarly edition projects, including the Vienna Wittgenstein Edition (1990–2001), Digital Variants (2004–2008), the Australian Electronic Scholarly Editions project (2012–2013), the Charles Harpur Critical Archive (2014-) and a pilot edition of Gianfrano Leopardi's Idilli (2018-). He currently works on developing practical web-based tools for making, visualising and publishing digital scholarly editions.

Colin Sippl (University of Regensburg, Germany – colin.sippl@ur.de) is currently a project employee at the University Library of Regensburg. Since 2017, he has been working on extending the open access services of the Electronic Journals Library (EZB). More recently, he has started developing and setting up a digital repository for literature, artefacts and experiments relating to the early life sciences based on the Invenio framework. He specialised in textual data mining and the development of media services in the institutional domain.

Elena Spadini (University of Lausanne - elena.spadini@unil.ch) is a postdoctoral researcher at the University of Lausanne. She holds a Ph.D. in Romance Philology from the University of Rome Sapienza (2016) and a M.A. in Digital Humanities from the École nationale des chartes (2014). She was a Marie Curie fellow in the IT Network DiXiT and co-directed the related volume Advances in Digital Scholarly Editing (Sidestone Press, 2017). She published in international journals and taught specialized courses in various European countries in the field of Digital Philology.

Francesca Tomasi (University of Bologna - francesca.tomasi@unibo.it) is associate professor in Archival Science, Bibliography and Librarianship at the University of Bologna (Italy). Her research is mostly devoted to digital cultural heritage, with a special attention to documentary digital edition, and a focus on knowledge organization methods in archives and libraries. She is member of different scientific committees of both associations and journals. In particular, she is President of the Library of the School of Humanities in the University of Bologna (BDU - Biblioteca di Discipline Umanistiche), Director of the international second cycle degree in Digital Humanities and Digital Knowledge (DHDK), President of the Italian Association of Digital Humanities (AIUCD – Associazione per l'Informatica Umanistica e la Cultura Digitale), and co-head of the Digital Humanities Advanced Research Center (/DH.ARC). She wrote about 100 papers and 4 monographs related to DH topics. She is editor and scientific director of several digital scholarly environments.

Athanasios Velios (University of the Arts London, United Kingdom – a.velios@arts.ac.uk) is Reader in Documentation at the University of the Arts London.

Georg Vogeler (University of Graz - georg.vogeler@uni-graz.at) is professor for Digital Humanities at the University of Graz and scientific director of the Austrian Center for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences. He is a trained historian (Historical Auxiliary Sciences). He spent several years in Italy (Lecce, Venice). In 2011, he became member of faculty at the Centre for Information Modelling at Graz University, where he was nominated full professor for Digital Humanities in 2016 and head of department in 2019. His research interests lie in late medieval and early modern administrative records, diplomatics (digital and non digital), digital scholarly editing and the history of Frederic II of Hohenstaufen (1194–1250). He was and is part in several national and international research projects related to his research interests.

Christian Wolff (University of Regensburg, Germany – christian.wolff@ur.de) has been Professor of Media Informatics at the Institute for Information and Media, Language and Culture at the University of Regensburg since 2003. He holds a PhD in information science and is a habilitated computer scientist. His research interests include: human-computer interaction, multimedia and web-based information systems, (multimedia) software engineering and information retrieval (in particular information literacy and social media).

Publications of the Institute for Documentology and Scholarly Editing / Schriftenreihe des Instituts für Dokumentologie und Editorik

01 Bernhard Assmann, Patrick Sahle. Digital ist besser. Die Monumenta Germaniae Historica mit den dMGH auf dem Weg in die Zukunft – eine Momentaufnahme. Norderstedt: Books on Demand, 2008. ISBN 978-3-8370-2987-1

02 Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age. Ed. by Malte Rehbein, Patrick Sahle and Torsten Schaßan in collaboration with Bernhard Assmann, Franz Fischer and Christiane Fritze. Norderstedt: Books on Demand, 2009. ISBN 978-3-8370-9842-6

03 Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2. Ed. by Franz Fischer, Christiane Fritze and Georg Vogeler in collaboration with Bernhard Assmann, Patrick Sahle und Malte Rehbein. Norderstedt: Books on Demand, 2010. ISBN 978-3-8423-5032-8

04 Birgit Jooss: Die digitale Edition der Matrikelbücher der Akademie der Bildenden Künste München. Ein Instrument zur Erforschung der Attraktivität einer international ausgerichteten Kunsthochschule (1808 – 1920). Norderstedt: Books on Demand, 2011. ISBN 978-3-8423-1278-4

05 Johannes Kepper: Musikedition im Zeichen neuer Medien – Historische Entwicklung und gegenwärtige Perspektiven musikalischer Gesamtausgaben. Norderstedt: Books on Demand, 2011. ISBN 978-3-8448-0076-0

06 Digitale Urkundenpräsentationen. Beiträge zum Workshop in München, 16. Juni 2010. Ed. by Georg Vogeler and Joachim Kemper. Norderstedt: Books on Demand, 2011. ISBN 978-3-8423-6184-3

07 Patrick Sahle: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe. Norderstedt: Books on Demand, 2013. ISBN 978-3-8482-6320-2

08 Patrick Sahle: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. Norderstedt: Books on Demand, 2013. ISBN 978-3-8482-5252-7

09 Patrick Sahle: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung. Norderstedt: Books on Demand, 2013. ISBN 978-3-8482-5357-9

10 Kodikologie und Paläographie im digitalen Zeitalter 3 / Codicology and Palaeography in the Digital Age 3. Ed. by Oliver Duntze, Torsten Schaßan and Georg Vogeler. Norderstedt: Books on Demand, 2015. ISBN 978-3-7347-9899-3

11 Kodikologie und Paläographie im digitalen Zeitalter 4 / Codicology and Palaeography in the Digital Age 4. Ed. by Hannah Busch, Franz Fischer and Patrick Sahle, in collaboration with Bernhard Assmann, Philipp Hegel, and Celia Krause. Norderstedt: Books on Demand, 2017. ISBN 978-3-7448-3877-1

12 Digital Scholarly Editions as Interfaces. Ed. by Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, and Gerlinde Schneider. Norderstedt: Books on Demand, 2018. ISBN 978-3-74810-925-9

13 Versioning Cultural Objects. Ed. by Roman Bleier and Sean M. Winslow. Norderstedt: Books on Demand, 2019. ISBN 978-3-7504-2702-0

14 Rekontextualisierung als Forschungsparadigma des Digitalen. Ed. by Simon Meier, Gabriel Viehhauser, and Patrick Sahle. Norderstedt: Books on Demand, 2020. ISBN 978-3-7519-1531-1

15 Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi, and Georg Vogeler. Norderstedt: Books on Demand, 2021. ISBN 978-3-7543-4369-2