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

Schriften des
Instituts für Dokumentologie und Editorik

herausgegeben von:

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

Projects and Editions

Appendices

Infrastructures and Technologies

Formal Models

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to encode entries which were originally published in printed form, but also incorporates amendments and additions from unpublished documents.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_cur rent_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_cur rent_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>

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).