Burrows, Toby (1); Fraas, Mitch (2); Hyvönen, Eero (3); Ikkala, Esko (3); Koho, Mikko (3); Lewis, David (1); Morrison, Andrew (1); Page, Kevin (1); Ransom, Lynn (2); Thomson, Emma (2); Tuominen, Jouni (3); Velios, Athanasios (4); Wijsman, Hanno (5)

Organisation(s): 1: University of Oxford, United Kingdom; 2: University of Pennsylvania; 3: Aalto University; 4: University of the Arts London; 5: Institut de recherche et d'histoire des textes

Long Paper Proposal

Linking Data to Explore the History of Medieval and Renaissance Manuscripts: the "Mapping Manuscript Migrations" Project

Abstract

Funded by the Digging into Data Challenge of the Trans-Atlantic Platform from 2017 to 2020, the "Mapping Manuscript Migrations" (MMM) project brings together data about the histories of more than 216,000 medieval and Renaissance manuscripts, for browsing, searching, and visualization. Four leading institutions from Great Britain, France, Finland, and the United States collaborated on this project, pooling their expertise in Semantic Web technologies and medieval manuscript curation and research, as well as contributing their own data from three contrasting datasets: the *Schoenberg Database of Manuscripts* at the University of Pennsylvania, the *Medieval Manuscripts Catalogue* at the University of Oxford, and the *Bibale* database from the Institut de recherche et d'histoire des textes (Paris). [1]

The datasets have been brought together in a Linked Open Data environment, with a unified data model based on the CIDOC-CRM and $FRBR_{OO}$ ontologies. [2] Public access to the aggregated data is available through the MMM semantic portal, using the "Sampo-UI" interface developed by the Semantic Computing Research Group at Aalto University in Helsinki. [3] The entire dataset, in the form of RDF Turtle files, can also be downloaded from the Zenodo data repository. [4] The scripts for transforming the data into RDF and the software for running the portal are also publicly available from GitHub. [5]

The proposed paper will explain how the Linked Open Data environment was constructed, beginning with the development of the MMM unified data model. This involved more than twelve months of weekly meetings by a project group which combined Semantic Web technical expertise with manuscript research expertise. The MMM data model was developed through an iterative process which drew on the data models used by the source datasets as well as a set of 24 typical research questions contributed by manuscript researchers. Transforming the source datasets into RDF triples and mapping them to the unified data model involved a distinct pipeline for each of the three sources. In the case of the Oxford *Medieval Manuscripts Catalogue*, this meant extracting relevant entities and relationships from more than 10,000 XML documents encoded with the Text Encoding Initiative's "Manuscript Description" guidelines, and transforming the resulting XML file into RDF triples.

The mapping process also involved matching key vocabularies present in the source datasets: manuscripts, persons, organizations, places, and works. For some of these, automatic reconciliation was possible through the use of common references to standard Linked Open Data vocabularies like VIAF (Virtual International Authority File) and TGN (Getty Thesaurus of Geographic Names). For others, more manual processes drawing on the expertise of manuscript researchers were required, such as matching the names of works on the basis of shared authors and similar titles, or matching manuscripts on the basis of a shared Phillipps number (for manuscripts formerly in the vast 19^{th}-century collection of Sir Thomas Phillipps).

The semantic portal developed by the project will also be presented and demonstrated. [6] The portal enables sophisticated browsing and searching across five main perspectives: Manuscripts, Works, Events, Actors (people and institutions), and Places. Using a variety of semantic filters, the data can be sifted and combined in ways that address a wide variety of complex research questions of this kind: "Find those manuscripts formerly owned by Alfred Chester Beatty which were produced in France in

the fifteenth century, which contain decorated initials, and which have a last known location in the United States." The interface also provides several map-based visualizations of the histories and movements of these manuscripts. Users can see the production places of the manuscripts on a world map and can zoom in to see specific towns and locations. A similar map is available for the last known locations of manuscripts, while a third map shows their migrations – in the form of an arc linking the place of production and the last known location.

The data for each entity can be viewed through a "landing page", which displays a table listing all the relationships attached to that entity. Users can also connect to the Linked Data Finland platform and see the same information expressed in the form of the properties contained in the data model. [7] The results sets from filtered browses and searches in the MMM portal interface can be exported as CSV spreadsheets through the Yasgui SPARQL query editor. The MMM dataset can also be queried directly through Yasgui, using the MMM SPARQL endpoint. [8] This makes it possible to construct complex and sophisticated SPARQL queries, taking full advantage of the richness of the MMM data model.

The RDF triples provided by the MMM project are, effectively, an additional layer over the source information. Users can always refer back to the original datasets via links provided in each entity's "landing page". They can also filter MMM data by source for direct access to a source's dataset if required. An unanticipated but welcome outcome has been the ability to help managers of the original datasets to identify problems. Because data correction is not part of the MMM transformation process, weaknesses, inconsistencies, and errors in the datasets become apparent in search results, alerting dataset managers that something needs to be fixed at their end. In this way, MMM enables managers to clean and enrich their data. For instance, in the SDBM and Bibale, hundreds of personal and institutional names have been corrected for authority control, resulting in a rich and as yet untapped record of names associated with manuscript production and trade. Additionally, more than 2,000 of the Oxford TEI files were updated by the MMM project with structured provenance information relating to previous collection owners, which can now be pulled into the RDF transformation.

To evaluate the portal, we used the same set of 24 research questions which informed the development of the data model. Each research question was run first against the native interfaces of the three contributing datasets. While each of these sources provides a relatively sophisticated interface, in almost every case it proved impossible to answer the questions fully. At best, the user was presented with a partial answer to the question, often in the form of a broader list of results which had to be scanned manually to identify relevant items. Some questions could not be answered at all using the source datasets alone (7 in Bibale, 7 in Oxford, 6 in Schoenberg). In the MMM portal, on the other hand, a majority of the questions (17 out of 24) could be answered readily and fully with a combination of filters and text searches. Only a few, more complex questions required further manual sifting of the result sets (7 out of 24). This group of questions was explored further by running queries against the MMM SPARQL endpoint.

The paper will examine the lessons learned from this kind of organizationally and culturally challenging international interdisciplinary collaboration, which has brought together manuscript researchers, librarians and curators, and computer science experts from four different countries. Other lessons learned in the course of the MMM project have included areas where future work would be valuable. One of these is the development of specialist Linked Open Data vocabularies for medieval studies. Manuscripts, in particular, need Linked Open Data identifiers if the data about them are to be linked effectively. The ISMI (International Standard Manuscript Identifier) initiative is working to define a manuscript identifier, but progress to date has been limited. [9]

Some re-thinking of the way in which provenance histories for manuscripts are recorded is also desirable. The goal should be to record and encode manuscript provenance data in a way which is sufficiently well-structured to map to a complex data model like that developed by MMM. As the MMM project demonstrates, this kind of approach makes it possible to construct a rich and innovative environment for asking and answering sophisticated questions about the history and provenance of medieval and Renaissance manuscripts.

References

1. https://sdbm.library.upenn.edu/ ; https://medieval.bodleian.ox.ac.uk/ ; http://bibale.irht.cnrs.fr/

2. http://cidoc-crm.org ; http://www.cidoc-crm.org/frbroo/home-0

3. https://github.com/SemanticComputing/sampo-ui See: Hyvönen, Eero, "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery," *Semantic Web*, vol. 11, no. 1 (2020), 187–193

4. https://doi.org/10.5281/zenodo.3667486

5. https://github.com/mapping-manuscript-migrations/

6. https://mappingmanuscriptmigrations.org

7. http://www.ldf.fi/dataset/mmm

8. http://ldf.fi/mmm/sparql

9. Cassin, Matthieu, "ISMI: International Standard Manuscript Identifier: Project of unique and stable identifiers for Manuscripts," *Manuscript Cataloguing in a Comparative Perspective: State of the Art, Common Challenges, Future Directions*, Centre for the Study of Manuscript Cultures, Hamburg, 7 - 10 May 2018:
https://www.manuscript-cultures.uni-hamburg.de/files/mss_cataloguing_2018/Cassin_pres.pdf