A New Model for Manuscript Provenance Research: The Mapping Manuscript Migrations Project

Toby Burrows
Oxford e-Research Centre, University of Oxford

Hanno Wijsman
Institut de la recherche et d'histoire des textes

Athanasios Velios
University of the Arts London

Jouni Tuominen
Semantic Computing Research Group, Aalto University

Emma Cawlfield Thomson
Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Lynn Ransom
Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Kevin Page
Oxford e-Research Centre, University of Oxford

Andrew Morrison
Bodleian Libraries, University of Oxford

David Lewis
Oxford e-Research Centre, University of Oxford

Mikko Koho
Semantic Computing Research Group, Aalto University

Esko Ikkala
Semantic Computing Research Group, Aalto University

Eero Hyvönen
Semantic Computing Research Group, Aalto University

Arthur Mitchell Fraas
University of Pennsylvania

Doug Emery
Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Following the obtention of a Round 4 Trans-Atlantic Platform Digging into Data Challenge grant in 2017, the Mapping Manuscript Migrations project (MMM) has been working to develop and test a methodology to link disparate data sets from Europe and North America with the aim of providing large-scale analysis and visualizations of the history and provenance of medieval and Renaissance manuscripts.[1] The work of the project has been carried out by four project partners: the University of Oxford (Oxford e-Research Centre and Bodleian Libraries), the Institut de recherche et d'histoire des textes, the University of Pennsylvania (Schoenberg Institute for Manuscript Studies), and Aalto University (Semantic Computing Research Group). Each partner was funded by its respective national funding agencies: the Economic and Social Research

1 Mapping Manuscript Migrations: http://blog.mappingmanuscriptmigrations.org/. For a report outlining the original aims of the project, see Toby Burrows, 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: A Journal of the Schoenberg Institute for Manuscript Studies* 3, no. 1 (2018): 249–52.

Council (United Kingdom), the Agence nationale de la recherche (France), the Institute of Museum and Library Services (United States), and the Academy of Finland.

Guided by a set of research questions identified at the outset of the project, MMM developed an innovative Linked Open Data model and data set that unifies three separate manuscript-related databases in a semantically consistent way, together with the workflows for transforming the institutional data contributions into the common structure. The data set has been made available through a Linked Open Data service hosted by the Linked Data Finland platform and the MMM semantic portal.[2] The latter was designed to test and demonstrate the platform for use by researchers. The portal, developed using the Sampo-UI framework, features faceted data search and exploration and ready-to-use visualization tools integrated with the user interface.[3] The aggregated data can thus be queried and visualized at scales ranging from a single manuscript to a total of more than 222,600 manuscripts as a group. Visualization tools developed in the portal show how the manuscripts have traveled across time and space from their place of production to their current locations, where they continue to find new audiences.

Although some concessions were made along the way, including limiting the number of data sets to just three and narrowing the scope of our research questions from general manuscript analysis to provenance-related analysis, the MMM project has been successful in demonstrating the possibilities of using Linked Open Data to enable large-scale analysis of the world's available manuscript data. The following report summarizes our methodology and results, and lays the groundwork for further research using our processes.

---

2 Eero Hyvönen, Esko Ikkala, Miho Koho, Jouni Touminen, Toby Burrows, Lynn Ransom, and Hanno Wijsman, "A Linked Open Data Service and Portal for Pre-modern Manuscript Research," *Proceedings of the Digital Humanities in the Nordic Countries 2019 Conference*, Copenhagen, March 2019, available at http://ceur-ws.org/Vol-2364/20_paper.pdf; Linked Data Finland: http://www.ldf.fi/dataset/mmm.

3 SemanticComputing: https://github.com/SemanticComputing/sampo-ui. See Eero Hyvönen, "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery," *Semantic Web* 11, no. 1 (2020): 187–93.

Data Modeling, Transformation, and Aggregation

MMM combines data from three specialist databases, each with a focus on the history and provenance of medieval and Renaissance manuscripts:

- *The Schoenberg Database of Manuscripts* (SDBM): a relational database containing more than 240,000 records for manuscript observations[4]
- *Bibale*: a relational database containing nearly 13,000 manuscript records[5]
- *Medieval Manuscripts in Oxford Libraries*: a collection of more than 10,000 XML documents[6]

Since each of these data sources is governed by its own customized data model, it was necessary to construct a unified data model that would be able to transform the underlying structures of the three data sets to allow them to function cooperatively in the MMM environment. This task was undertaken by the MMM Modelling Group, consisting of the team's Linked Data and data management specialists and a librarian. Working over a period of eighteen months, the Modelling Group's first tasks were to identify the requirements of manuscript provenance researchers, analyze the data models in the source data sets, and compare them with the International Council for Documentation Conceptual Reference Model (CIDOC-CRM) and Functional Requirements for Bibliographic Records-object oriented ($FRBR_{OO}$) ontologies.[7] Applying the MMM data model to the transformation pipeline proved to be more complicated than initially expected, but the iterative, trial-and-error process of applying the model revealed refinements throughout the process that would better support use and discovery in the final

4 A project of the Schoenberg Institute for Manuscript Studies, University of Pennsylvania Libraries: https://sdbm.library.upenn.edu/.

5 A project of the Institut de recherche et d'histoire des textes, Centre national de recherche scientifique: http://bibale.irht.cnrs.fr/.

6 A project of the Bodleian Libraries, University of Oxford: https://medieval.bodleian.ox.ac.uk/.

7 CIDOC-CRM: http://cidoc-crm.org; $FRBR_{OO}$: http://www.cidoc-crm.org/frbroo/home-0.

product. Final adjustments to the model were still being identified during the implementation and testing of the MMM portal interface to the data. Drawing on the CIDOC-CRM and $FRBR_{OO}$ ontologies, with the addition of entity classes and relationships specific to MMM, the unified data model successfully combines the three data sets. The unified model is ontologically based on five main classes of entities identified in the development of the modeling process. These entities represent five points of topic alignment in each of the data sets: Manuscripts, Works, Actors, Places, and Events. As part of the transformation, individual Actors and Places occurring in each data set were matched and reconciled using the Virtual International Authority File (VIAF) and Getty Thesaurus of Geographic Names (TGN) Linked Open Data authorities.[8] Works and Manuscripts that were duplicated across two or more data sets were matched and reconciled through the manual comparison of specific entities identified by string similarity.[9]

Once the design of the unified data model was complete, the original data from the three sources were transformed into Resource Description Framework (RDF) triples—sets of three entities that together yield a statement about semantic data—and mapped to the MMM data model.[10] The process for converting the Text Encoding Initiative (TEI) Extensible Markup Language (XML) documents that comprise the data for the *Medieval Manuscripts in Oxford Libraries* catalog involved an additional set of preparatory scripts as well.[11] In this case, the initial step was to extract a

8 VIAF: http://viaf.org/; Getty TGN: http://www.getty.edu/research/tools/vocabularies/tgn/index.html.

9 Toby Burrows, Antoine Brix, Douglas Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Synnove Myking, Lynn Ransom, Emma Cawlfield Thomson, Jouni Tuominen, Hanno Wijsman, and Pip Wilcox, "Linked Open Data Vocabularies and Identifiers for Medieval Studies," *Proceedings of Digital Humanities in Nordic Countries (DHN 2020), Riga*, CEUR Workshop Proceedings, March 2020, available at https://seco.cs.aalto.fi/publications/2020/burrows-et-al-dhn-2020.pdf.

10 RDF triples: https://www.w3.org/RDF/. Scripts and documentation for the data conversion pipeline are available on GitHub: https://github.com/mapping-manuscript-migrations/mmm-data-conversion.

11 TEI: https://tei-c.org/. For the preparatory scripts, see https://github.com/mapping-manuscript-migrations/bodleian-mmm.

selection of TEI tags from each of these documents and assemble these into a single XML file.

The RDF triples provided by MMM are, effectively, a supplementary layer to the source information. Users can always refer back to the original data sets via links provided in each entity's landing page. They can also filter MMM data by source for direct access to a source's data set if required. The transparent relationship between the source data sets and the MMM data underscores the role of MMM as an aggregator of data rather than as a data management system, though an unanticipated but welcome outcome has been its ability to help managers of the original data sets to identify problems. Because data correction is not part of the MMM transformation process, weaknesses, inconsistencies, and errors in the data sets become baldly apparent in search results, alerting data set managers to the need for a fix 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 two thousand of the Oxford TEI files have been updated with structured provenance information relating to previous collection owners that can now be pulled into the RDF transformation.

Data Services Online

The MMM Linked Data are served by the Linked Data Finland platform.[12] There are more than twenty million RDF triples in the MMM data set. A SPARQL endpoint enables all the RDF triples to be searched directly using the SPARQL query language.[13]

In addition to SPARQL queries, the data service supports the best practices of the World Wide Web Consortium (W3C) for publishing Linked Data, including the following types of data access mechanisms:[14]

---

12 Linked Data Finland: http://www.ldf.fi/dataset/mmm.

13 Available at http://ldf.fi/mmm/sparql.

14 Tom Heath and Christian Bizer, *Linked Data: Evolving the Web into a Global Data Space* (San Rafael, CA: Morgan & Claypool, 2011).

- Resolving Uniform Resource Identifiers (URIs)
- Viewing the RDF description of a URI
- Linked Data browsing starting from a URI
- Downloading the data

If this URI/URL is used in a browser, a human-readable view of the data resource is rendered. By changing the HTTP request's "Accept" header, the underlying RDF data can be accessed, too.

MMM Portal

The MMM semantic portal was implemented using the Sampo-UI framework developed by the Semantic Computing Research Group (SeCo) at Aalto University and the Helsinki Centre for Digital Humanities (HELDIG).[15] It works by constructing SPARQL queries against the Linked Data service and presenting the results in a user interface that allows for browsing and searching. Five basic perspectives on the data are available, based on the main classes of entities: Manuscripts, Works, Events, Actors, and Places (fig. 1).

In each of these views, a set of filters can be used to refine and explore the data. For Manuscripts, these include: author, work, language, place of production and date, owner, collection, last known location, and decoration (fig. 2). Each individual entity has a landing page, which provides all the available data, together with a link to the Linked Data resource for that item.

Visualizations of the data are also available. For the Manuscripts perspective, these include places of production plotted on a world map (fig. 3). Last known locations can be shown on a separate map. Zooming in on the maps shows a list of manuscripts associated with specific places. A visualization, which shows the arc linking a place of production of a given manuscript to its last known location, is also available for migrations.

---

15 Sampo-UI framework: https://github.com/SemanticComputing/sampo-ui; Semantic Computing Research Group: http://seco.cs.aalto.fi/; Helsinki Centre for Digital Humanities: http://heldig.fi.

FIGURE 1. The home page for the MMM semantic portal, with icons directing users to the five data perspectives.

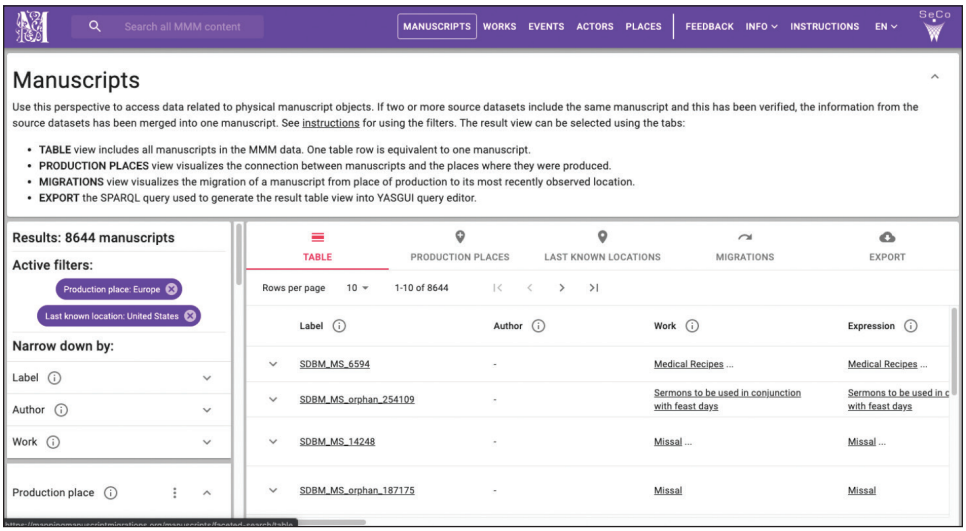

FIGURE 2. Tabular search results from the Manuscripts perspective, returning data for manuscripts produced in Europe with last known locations in the United States.

FIGURE 3. The places of production for manuscripts in the MMM data set. The numbers within each marker indicate the number of manuscripts produced in that region or location. Some coordinates are approximate.

General trends in the migrations of manuscripts—especially from Europe to North America—can be seen at the broadest level of this visualization (fig. 4). Places associated with Events and Actors can also be seen on a world map. All of these visualizations are dynamic. They start with the entire result set (e.g., all Manuscripts) and change as the user refines the result set through the use of the filters.

The results of queries can also be exported from the portal in the form of a Comma-separated values (CSV) spreadsheet. Each result set includes a link to Yasgui, which is a public SPARQL query service.[16] Yasgui provides a tabular view of the result set, which can then be downloaded as a CSV, XML, or JavaScript Object Notation (JSON) file (fig. 5).

16 Yasgui: https://yasgui.triply.cc/.

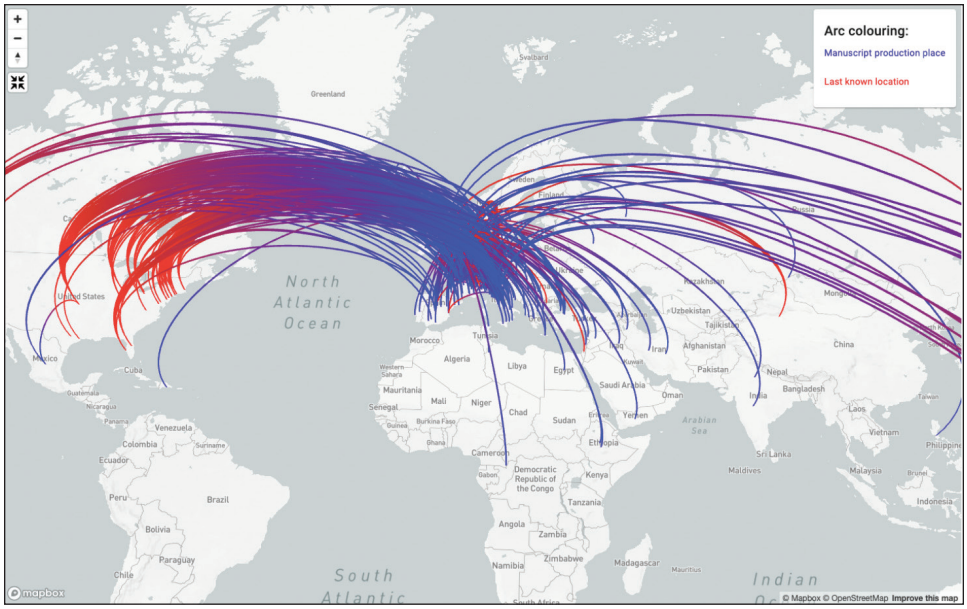

FIGURE 4. Migration visualization of the 8,747 manuscripts in the MMM data set that were once owned by Sir Thomas Phillipps, from their places of production to their last known locations.

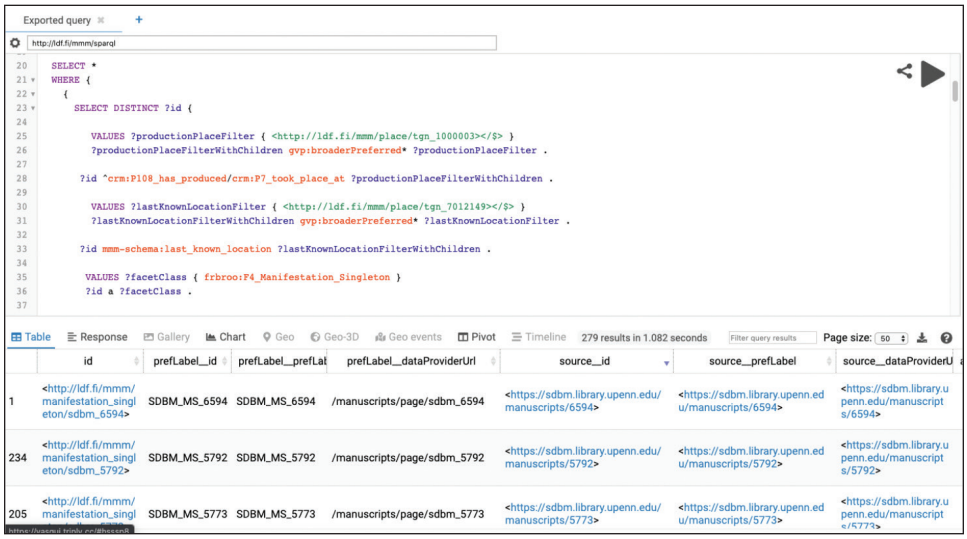

FIGURE 5. The result set from figure 2 as it appears in the Yasgui query service, where users can download the data.

Data Reuse

The MMM project is committed to open access and sharing our data and research results. A copy of the MMM aggregated data has been deposited in the Zenodo data repository. Version 2.1.0 of the data—amounting to about 1.3 gigabytes in total—was deposited on 8 September 2020.[17] The data are made available as RDF Turtle files.[18] There is one file for each of the three source data sets, containing the transformed and mapped source data in the form of RDF triples, together with the reconciled instances of Manuscripts, Works, and Actors. Also deposited are a separate Places file, which contains the RDF triples for the reconciled Places, and a Schema file. The Schema file contains the unified data model used for the MMM data.[19]

Other ways of reusing the data are also being supported. A Docker container for mounting the RDF data into a Fuseki triple store has been made available on GitHub.[20] The Oxford e-Research Centre is experimenting with pointing to a different user interface at the MMM SPARQL endpoint, using the ResearchSpace software developed by the British Museum and metaphacts GmbH.[21]

Evaluation

The project team developed a set of twenty-four research questions to guide its progress. Some of these questions were elicited from an initial focus group of manuscript researchers, held at Oxford University in 2017. Others were contributed by members of the MMM project team or taken from a

17 Available at https://doi.org/10.5281/zenodo.4019643.

18 See https://www.w3.org/TR/turtle/.

19 See http://ldf.fi/schema/mmm.

20 See https://github.com/mapping-manuscript-migrations/mmm-fuseki.

21 Dominic Oldman 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. Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa et al. (Berlin: Springer, 2018), 325–40.

list produced by the French project Biblissima.[22] Some of these questions were specific—for example: "Which manuscripts containing texts by Ramon Llull were sold in the nineteenth century?" Others were more generic, such as: "How many illuminated manuscripts were in a particular collection?"

These research questions were used in the evaluation of the MMM portal and its Linked Open Data framework. Each question was tested first against the three source data sets individually. While each of these sources provides a relatively sophisticated interface, in almost every case it proved difficult to answer the questions fully (Table 1). At best, the user was presented with a partial answer to the question, often in the form of a broader list of results that had to be scanned manually to identify relevant items. Some questions could not be answered at all using the source databases alone (seven in Bibale, seven in *Medieval Manuscripts in Oxford Libraries*, and six in the SDBM).

In the MMM portal, on the other hand, a majority of the questions (seventeen out of twenty-four) could be answered readily with a combination of filters and text searches. Only a few of the more complex questions required further manual scanning of the result sets (seven out of twenty-four). This group of questions was explored further by running queries against the MMM SPARQL endpoint.

Future Work

The MMM project has demonstrated the value and effectiveness of a Linked Open Data approach to aggregating provenance data for medieval and Renaissance manuscripts. The sophisticated data model and the transformed data have been made available for reuse, while the MMM portal shows how searching, filtering, and visualizing can be successfully applied to answer complex research questions. An obvious desideratum for the future is to transform and incorporate data from a wider range of sources, with the aim

22 Biblissima, "Ontologie Biblissima: Méthodologie: Requêtes intéressantes (vision/proposition)," available at https://doc.biblissima.fr/ontologie-biblissima.

of both increasing the number of manuscripts covered and adding to the information available for the 222,600 manuscripts currently represented.

The project has also identified areas where future work would be valuable. One of these is the development of specialist Linked Open Data vocabularies for medieval studies. MMM was able to make effective use of generic vocabularies such as VIAF for Actors and Getty TGN for Places, drawing on their use in the source data sets. This is largely because the history of manuscripts actually focuses on the nineteenth and twentieth centuries as much as on the medieval period, which reduces the need to identify medieval names. Nevertheless, there are substantial numbers of medieval people (as authors and manuscript owners), medieval organizations (especially religious houses), medieval works, and even medieval places that do not appear in the more generic vocabularies. There are many extant lists and databases of medieval names that could be transformed into Linked Open Data, which would be of great benefit to discovery services and knowledge graphs in this field of research.

Manuscripts, in particular, need Linked Open Data identifiers if the data about them are to be linked effectively. The current default approach of identifying a manuscript by its owner, collection, and shelfmark raises a number of problems, including common inconsistencies in the formatting of shelfmarks, even within the same institution or collection. In addition, frequent changes in ownership for some manuscripts and the fact that institutional owners often change their exact nomenclature mean that more precision in manuscript identification is needed. The MMM project used Phillipps numbers as one way of reconciling data about the same manuscript, which matched nearly nine thousand manuscripts. The International Standard Manuscript Identifier (ISMI) initiative is working to define a universal manuscript identifier, but progress to date has been limited.[23]

Some rethinking of the way in which provenance histories for manuscripts are recorded is also desirable. The MMM project worked with two specialized provenance-oriented databases (Bibale and the Schoenberg

23 François Bougard et al., "International Standard Manuscript Identifier (ISMI): pour un registre électronique des identifiants des livres manuscrits," *DigItalia* (2020/1): 45–52; available at http://digitalia.sbn.it/article/view/2486

Database of Manuscripts), and this provided a substantial amount of well-structured data, even if these databases use quite different data models. Records for manuscripts in library MARC-based catalogs, in contrast, usually record provenance information in a note that is not structured at all. The MMM project did not attempt to incorporate such data, which would have required the use of text analysis and entity recognition techniques.

The Bodleian Library's TEI-XML documents, on the other hand, rely heavily on the <provenance> tag to encode information about manuscript ownership. This tag is geared toward the kinds of narrative histories and recording of provenance evidence found in traditional printed manuscript catalogs. While the Bodleian has also encoded the names of persons mentioned in these narratives, it proved difficult to extract anything more than a generic event from this kind of data. A possible way forward for TEI-encoded catalogs might be found in the work of the Linked Art project, which is developing a model for provenance data within its more general schema for art history, using CIDOC-CRM.[24]

The goal should be to record and encode manuscript provenance data in a way that 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.

24 Linked Art, "Linked Art Data Model," available at https://linked.art/model/index.html.