Mapping Manuscript Migrations Knowledge Graph: Data for Tracing the History and Provenance of Medieval and Renaissance Manuscripts

Toby Burrows, Douglas Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Andrew Morrison, Kevin Page, Lynn Ransom, Emma Cawlfield Thomson, Jouni Tuominen, Athanasios Velios, and Hanno Wijsman

Introduction

The Mapping Manuscript Migrations project (MMM) links disparate datasets from Europe and North America to provide an international view of the history and provenance of medieval and Renaissance manuscripts.[1] The aggregated data can be browsed and visualized at scales ranging from an individual manuscript to more than 216,000 manuscripts in total. The tools developed 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.

MMM has two components. The first, which is the focus of this paper, consists of a Linked Open Data service hosted by the LDF.fi platform: http://www.ldf.fi/dataset/mmm The second component is the MMM semantic portal, which is designed to test and demonstrate the platform for use by researchers: https://mappingmanuscriptmigrations.org/ It includes, in addition to faceted data search and exploration, a variety of ready-to-use Digital Humanities tools integrated seamlessly with the user interface. The MMM portal was implemented using the Sampo-UI framework.[2]

Data Transformation and Aggregation

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

- *Schoenberg Database of Manuscripts*: https://sdbm.library.upenn.edu/ (a relational database containing more than 240,000 records for manuscript observations)
- *Bibale*: http://bibale.irht.cnrs.fr/ (a relational database containing nearly 13,000 manuscript records)
- *Medieval Manuscripts in Oxford Libraries*: https://medieval.bodleian.ox.ac.uk/ (a collection of more than 10,000 XML documents)

The data have been aggregated using a set of shared ontologies and a novel unified Data Model that extends the CIDOC-CRM and $FRBR_{OO}$ ontologies. Instances of the four main classes of entities (Manuscripts, Works, Actors, and Places) have been reconciled in two ways: automatically through the use of Linked Open Data authorities like VIAF (Virtual International Authority File) and TGN (Thesaurus of Geographic Names) where possible, as well as by manual comparison of specific entities identified by string similarity.[3]

---

[1] Toby Burrows, Eero Hyvönen, Lynn Ransom, 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*, vol. 3, no. 1 (2018), 249-252. link

[2] Eero Hyvönen, "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery," *Semantic Web Journal*, 2020. In press. pdf

[3] 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. pdf

The original data have been transformed into RDF triples and mapped to the MMM Data Model. Scripts and documentation for the MMM data conversion pipeline are available on GitHub. The process for converting into RDF the Text Encoding Initiative (TEI) XML documents which comprise the data for the *Medieval Manuscript in Oxford Libraries* catalogue involves an additional set of preparatory scripts as well. In this case, the initial step is to extract a selection of TEI tags from each of these documents and assemble these into a single XML file.

The original data have not been corrected or amended in any way by the MMM project. The source information for each resource in the unified data has been retained by MMM, so that users can always refer back to the original dataset and can limit their use of the MMM data by source if required. Errors and omissions in the data should be reported to the owners of the source datasets.

Zenodo Data Deposit

A copy of the MMM aggregated data has been deposited in the Zenodo data repository. Version 1.1.0 (14 February 2020) of the data – amounting to about 1.25 GB in total – is available here: https://zenodo.org/record/3667486

The data are made available as RDF Turtle files. There is one file for each of the three source datasets, containing the transformed and mapped source data in the form of RDF triples, and including 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. Documentation about the schema is available here: documentation. As well as some MMM-specific classes and properties, the MMM schema makes use of the following vocabularies:

- RDF: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- RDFS: http://www.w3.org/2000/01/rdf-schema#
- Erlangen CRM: http://erlangen-crm.org/current/
- Erlangen FRBRoo: http://erlangen-crm.org/efrbroo/
- Getty Vocabulary Program ontology: http://vocab.getty.edu/ontology#
- SKOS: http://www.w3.org/2004/02/skos/core#

Data Services Online

The linked data are served by the Linked Data Finland Linked Open Data service, hosted at: http://www.ldf.fi/dataset/mmm/

For searching and reusing all the underlying data using the SPARQL query language, the SPARQL endpoint is available at: http://ldf.fi/mmm/sparql

In addition to SPARQL queries, the data service supports the following types of data access mechanisms:

- Viewing the RDF description of a URI;
- Linked Data browsing starting from a URI.

A typical example of a URI for an MMM resource can be seen here: http://ldf.fi/mmm/manifestation_singleton/sdbm_784

Data Reuse

The MMM data are made available for reuse under a CC BY-NC 4.0 license. Two main reuse cases are envisaged, both of which would be applicable to researchers studying such some subjects as the history of medieval and Renaissance manuscripts, the history of collecting and collections, and the transmission and dissemination of classical, medieval, and Renaissance texts. The first case would cover the whole dataset; there have already been sixteen downloads from the Zenodo repository in the first two months of availability. The Oxford e-Research Centre has loaded a copy of the entire dataset into a different software environment – *ResearchSpace* (developed by MetaPhacts and the British Museum) – and is currently configuring a new interface, which will include a network visualization of the data.[4] The second case applies to a selection of the data, identified through the portal or a SPARQL query. One of the authors (Burrows) is downloading a sub-set of the data relating to a specific manuscript collector (Sir Thomas Phillipps) for import into a *nodegoat* database of Phillipps manuscripts, using CSV spreadsheets as the transport mechanism.[5]

The MMM dataset also provides a series of reusable Linked Open Data vocabularies for manuscripts, actors (persons and organizations), works, and places. Each entity is published with a URI which meets LOD standards, and with cross-references to other widely-used LOD vocabularies for these types of entities, where relevant. This is particularly valuable for those entities which do not have identifiers in a generic vocabulary like VIAF, Wikidata, Library of Congress, Bibliothèque nationale de France, or others. There are more than 23,100 actors (43%) and 470 places (10%) without such identifiers. For manuscripts, MMM offers the first dataset which creates a LOD identifier for a large number of manuscripts (more than 217,700) and matches it to their institutional shelf-mark where applicable. These vocabularies will be of significant value to future efforts to build Linked Open Data services for medieval and Renaissance studies.

Acknowledgments

The Mapping Manuscript Migrations project was funded under Round 4 of the Trans-Atlantic Platform's Digging into Data Challenge. The four project partners are 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 their respective national funding agencies: Economic and Social Research Council (UK), Agence nationale de la recherche (France), Institute of Museum and Library Services (US), and the Academy of Finland.

[4] 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), pp. 325-340.

[5] Toby Burrows, "The History and Provenance of Manuscripts in the Collection of Sir Thomas Phillipps: New Approaches to Digital Representation," *Speculum* 92 S1 (Oct. 2017), S39-S64