A Linked Open Data Service and Portal for Pre-modern Manuscript Research

Eero Hyvönen[1,2], Esko Ikkala[1], Jouni Tuominen[1,2], Mikko Koho[1], Toby Burrows[3], Lynn Ransom[4], and Hanno Wijsman[5]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
[2] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
[3] Oxford e-Research Centre, University of Oxford, UK
[4] Schoenberg Institute for Manuscript Studies, University of Pennsylvania, USA
[5] Institut de recherche et d'histoire des textes, France

Abstract. This paper presents a Linked Open Data publishing model for aggregating data from heterogeneous, distributed pre-modern manuscript databases into a global, harmonized data model and service. Our research hypothesis is that on top of the global data service based on ontologies and well-defined semantics, tools and applications can be created for solving novel research problems in manuscript studies using Digital Humanities methods. First results in implementing such a system in the international Mapping Manuscript Migrations project are described with lessons learned discussed in dealing with complex and imperfect historical data.

1 Introduction

Thousands of European pre-modern manuscripts have survived until the present day. As the primary surviving witnesses to the world of pre-modern Europe they provide crucial evidence for research in many disciplines, including textual and literary studies, history, cultural heritage, and the fine arts. [3] As the result of changes of ownership over the centuries, they are now spread all over the world. They often feature among the treasures of libraries, museums, galleries, and archives, and they are frequently the focus of exhibitions and events in these institutions.

Over the last twenty years there has been a proliferation of digital data relating to these manuscripts, including catalogues, specialist databases, and numerous collections of digital images – many of them IIIF-compliant[6]. But there is little in the way of coherent, interoperable digital infrastructure, with the result that large-scale discovery and analysis requires the time-consuming exploration of numerous disparate resources.

This paper introduces the Mapping Manuscript Migrations (MMM) project[7] [2] that aims to address this problem, and presents its first experiences and results from a technical, Linked Data publishing perspective. Our goal is to build a Linked Open Data (LOD) [5,6] framework and web system for harmonizing manuscript data from various

[6] https://iiif.io
[7] http://mappingmanuscriptmigrations.org

disparate sources, in order to provide researchers with searchable and browsable access to aggregated evidence about the history of pre-modern manuscripts.

In the following, we first present the publications model underlying the MMM initiative, where data from distributed heterogeneous manuscript data silos is converted and aggregated into a harmonized form and published as the *MMM Data Service*, based on Linked Data principles and best practices of W3C[8]. After this, a workbench *MMM Portal* under construction for studying the manuscripts is described. In conclusion, lessons learned from the work thus far are summarized, and directions for further research are outlined.

2 Publishing and Harmonizing Manuscript Metadata

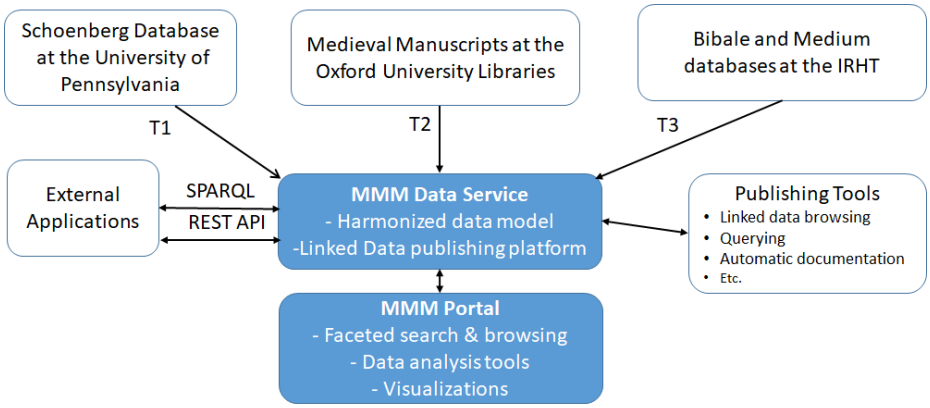

Fig. 1. MMM publishing model

Publishing Model Figure 1 depicts the publication model of the MMM system. In its initial phase, the project brings together more than 450 000 records from four important European and North American datasets, which use very different data models: the Schoenberg Database of Manuscripts[9], Medieval Manuscripts in Oxford Libraries[10], and Bibale[11] and Medium[12] [11] of the Institut de recherche et d'histoire des textes (IRHT). These data are transformed (T1–T3 in the Figure) into a unified harmonizing data model used in the MMM Data Service platform that is depicted in the middle of the Figure.

[8] https://www.w3.org/standards/techs/linkeddata#w3c_all
[9] https://sdbm.library.upenn.edu
[10] https://medieval.bodleian.ox.ac.uk
[11] http://bibale.irht.cnrs.fr
[12] http://medium.irht.cnrs.fr

Harmonizing Data Model As each of the source manuscript datasets involved in the project have their own preconditions and goals, and thus follow their own data modeling conventions, a unified data model for harmonizing them is needed. This model – still under some development and to be published in detail later on – is based on input from manuscript researchers with the goal of being semantically sufficient for answering a wide range of research questions. The most essential core classes in the model are 1) *physical manuscript objects* with properties such as *shelf-mark*, *owner* and *title*, and 2) *provenance events* that describe events related to the manuscripts, such as production, appearance in an auction, acquisition, etc. The model makes use of the the CIDOC CRM[13] [4] and FRBRoo [9] ontologies, where the essential distinctions between *works*, *expressions*, *manifestations*, and *items* are considered [8]. These ontology models were chosen as the basis because they are modern standards for data harmonization in museums and libraries, respectively, and they support event-based modeling needed in modeling provenance data that are essentially chains of events concerning the manuscript objects.

The MMM publishing model facilitates the aggregation and concurrent use of heterogeneous, distributed datasets with shared user interfaces, tools, and SPARQL queries. In our work, the MMM Portal is implemented on top of the MMM Data Service (the bottom in the Figure 1), but also other applications (left in the Figure) can make use of the data in a similar way, including also the original data providing web services. The MMM Data Service also includes a wide range of tools for managing and using linked data, such as a linked data browser, SPARQL query interfaces, automatic documentation tool, and so on. These services are provided by the Linked Data Finland platform[14] that is used as the basis [7] for the MMM Data Service.

Data Transformation The data harmonization work was initiated by investigating the source datasets, starting with the Schoenberg Database. In our first demo system, the MMM Data Service contains a version of this dataset that is mapped into the harmonizing data model. Other datasets and transformations will be integrated later on in the system. Integrating the datasets has involved building individual pipelines for transforming the source datasets into simple RDF formats by the data providers. Three of them are customized relational databases, while the fourth dataset – the Bodleian Library's catalogue – consists of XML documents encoded in accordance with the Text Encoding Initiative (TEI) Manuscript Description guidelines[15]. A special pipeline has been built by the data provider to extract and convert a selection of elements from these TEI documents into a record-like form more suitable for transformation to RDF. In the case of the Schoenberg Database, a new SPARQL endpoint has also been implemented by the data provider, which is available for general use[16].

Matching between the various datasets has initially focused on shared places and persons. Two of the four data sources have now annotated their records with identifiers from the Getty Thesaurus of Geographic Names (TGN)[17] and three with those from

[13] http://cidoc-crm.org
[14] http://ldf.fi
[15] http://www.tei-c.org
[16] https://sdbm.library.upenn.edu/sparql-space
[17] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

the Virtual International Authority File (VIAF)[18], as well as references to a range of other vocabularies. These identifiers have been used to match places and persons in the aggregated MMM project data. Matching manuscripts themselves is more problematic, since there is currently no standard for constructing and managing unique identifiers for manuscripts, though an International Standard Manuscript Identifier has been under discussion[19]. MMM has been testing and evaluating different approaches for assigning identifiers, including the use of ARKs (Archival Resource Keys)[20] from the Bodleian Library and in the IRHT's Medium database.

3 MMM Portal for Manuscript Studies

The goal of the MMM Portal is to provide a search and discovery interface for users with or without clearly defined research questions. The portal offers four main application perspectives based on the following classes of aggregated MMM project data: 1) Manuscripts, 2) Places, 3) People, and 4) Organizations. The instances of the core classes can be presented to the user as a paginated table, on a map based on various geographical information of the instances, and as a percentage frequency distribution based on arbitrary properties of the instances.

In each application perspective, the focus is on enabling the user to both explore and browse the data freely and identify a group of instances of the core classes based on a combination of criteria. In the Manuscripts perspective, a combination of criteria could be manuscripts *produced in Castile, including Spanish texts, previously owned by English private collectors, currently owned by an institution in North America.* Faceted search [10] is an effective paradigm for formulating such criteria in a user-friendly way.

At the moment a first version of the Manuscripts application perspective has been implemented. The perspectives for places, people and organizations can be constructed in a similar fashion by re-using the components of the Manuscripts perspective. Figure 2 depicts the Manuscripts perspective with faceted search. By default all manuscripts are shown in a paginated table. The key properties of manuscripts, such as shelf-mark, author or creation date, can be filtered with facet selections. Figure 3 shows an example of a hierarchical facet selection with search functionality. Besides hierarchy, the facet selections need to support further filtering. For example, the author facet must provide an additional filter for the birth date of the authors. All facet selections are connected, so whenever the user makes a selection, the value list of other facets is updated. This way it is impossible to end up with an empty result set by using any combination of the facets.

Moreover, each instance is associated with an information "home page" with an aggregated description on the instance and how it is related to other instances. For a person instance there could be, e.g., lists of related manuscripts based on different roles such as author, scribe or owner.

[18] https://viaf.org

[19] https://www.irht.cnrs.fr/?q=fr/agenda/manuscript-ids-identifiants-des-manuscrits

[20] https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8793

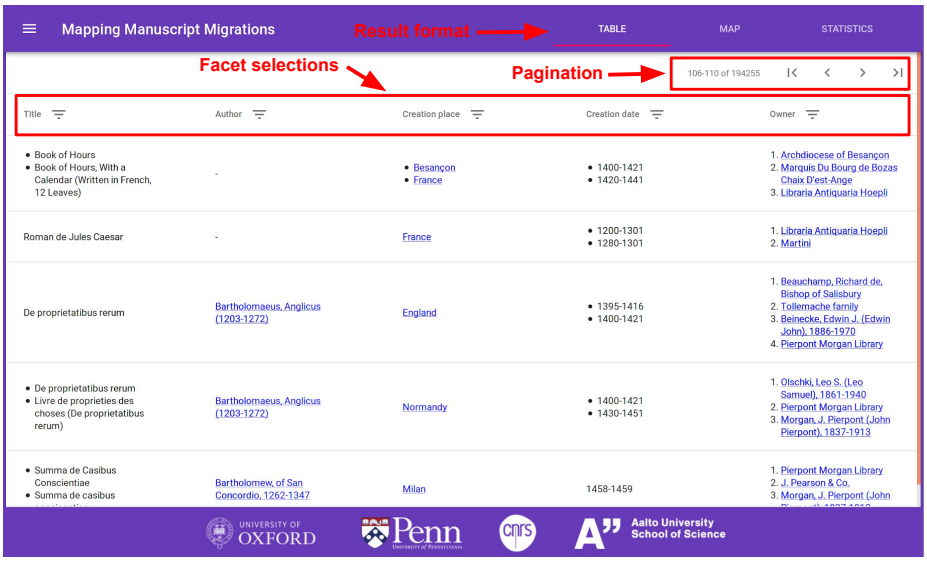

Fig. 2. MMM Portal: Faceted search and browsing of manuscripts in tabular view

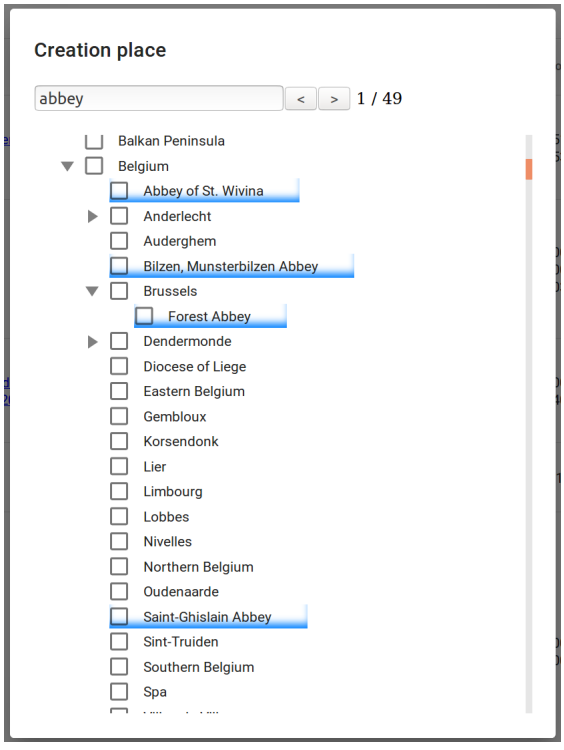

Fig. 3. MMM Portal: Hierarchical facet with search functionality for manuscript creation places

Figure 4 illustrates how the aggregated manuscript data (with optional filters selected by the user) are rendered on a map with clustered map markers, based on the creation places of the manuscripts. The numbers on clusters and markers indicate how many manuscripts were created in the area or specific place in question. When zooming in closer, Figure 5 shows how historical map sheets aligned on modern maps can be used to provide contextual information. However, using historical maps in this way is problematic in many ways and further user interface research is needed in order not give the end user wrong impressions about the data. Here 5 dots are spread over 5 spots in Paris, but there are also lots of dots that are all in one place corresponding to the general annotation ”Paris” that occurs frequently in the data. In general the place annotations can anything between a continent and a specific building, so a method for visualizing the varying granularity of geocoded data is obviously needed.

Furthermore, the times of the maps shown typically do not match with the times of the underlying manuscripts (or other data) that are also usually different from each other, too. For example, the place selected in Figure 5 is in fact the geographical spot of the current building of the Bibliothèque nationale de France (BnF), which did not exist at the time of the map (neither the BnF as an institution, nor the building). The predecessor of the BnF, the French Royal Library, moved to this spot in the 18th century (well after 1705), but the urban landscape as depicted on the 1705 map has changed since.

Fig. 4. MMM Portal: Faceted search and browsing of manuscripts in global map view

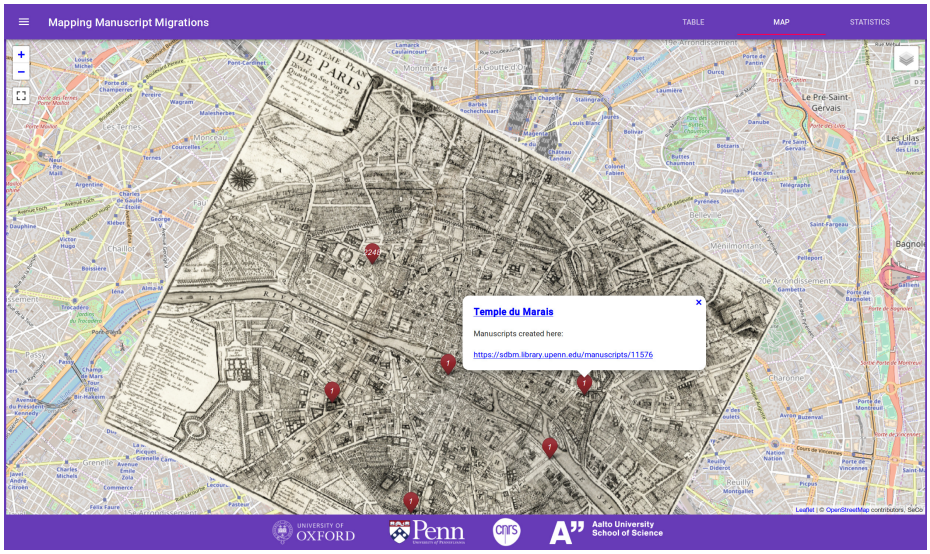

Fig. 5. MMM Portal: Manuscript creation places on OpenStreetMap base layer with a semi-transparent map of Paris in 1705 layered on top.

The general architecture of the MMM Portal[21] is presented in Figure 6. The system consists of a NodeJS[22] backend build with Express framework[23] (in the middle) and a frontend based on React[24] and Redux[25] (on the right). The MMM Data Service is shown on the left. An instance[26] of MapWarper[27] (on the left) can be used for aligning and publishing historical maps. When designing the architecture, the main goal of the backend was to ease the combining of attribute data from multiple SPARQL endpoints and raster data from various spatial data sources into a React frontend.

The data is published on the Linked Data Finland platform, which is powered by a combination of Fuseki SPARQL servers[28] running in Docker containers[29] for storing the primary data and a Varnish Cache web application accelerator[30] for routing URIs and content negotiation.

[21] https://github.com/SemanticComputing/mmm-web-app
[22] https://nodejs.org/en/
[23] https://expressjs.com
[24] https://reactjs.org
[25] https://redux.js.org
[26] http://mapwarper.onki.fi
[27] https://mapwarper.net
[28] https://jena.apache.org/documentation/fuseki2/
[29] https://www.docker.com
[30] https://varnish-cache.org

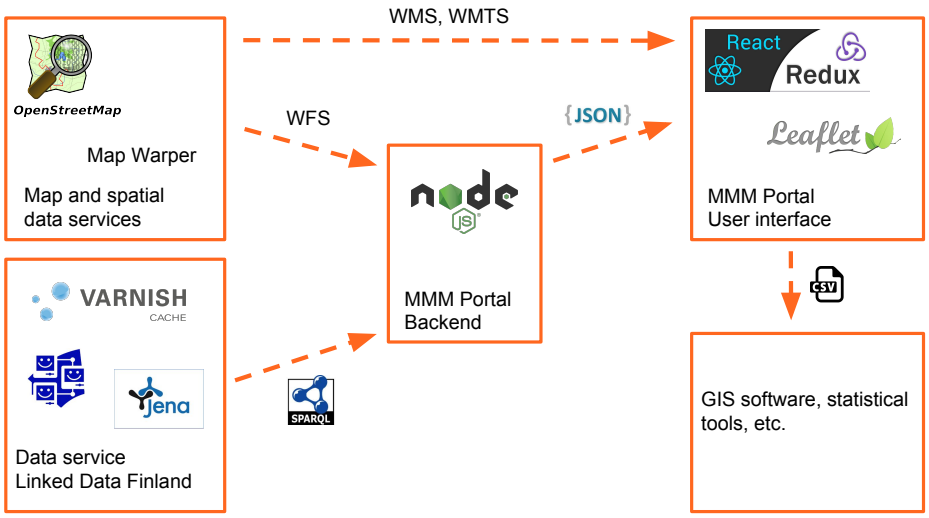

Fig. 6. MMM Portal architecture

4 Lessons Learned

Digging into manuscript data has turned out in many ways more challenging from a data modeling and technical perspective than expected. Defining the very concept of ”the manuscript” itself raised many ontological modeling questions, since manuscripts can be just fragments of a whole, can be separated into parts, copied, annotated, and united to others over time. Also identifying records describing the same manuscript can be very hard, in many cases probably impossible, as they have been described in different contexts in different ways, in terms of different titles, and in different languages. There is no unique identifier scheme for manuscripts, in contrast to printed books, and library shelf-marks are not quoted consistently or accurately. The amount of data is also fairly large, hundreds of thousands of records, which sets efficiency requirements for the technical solutions.

The data are often also incomplete, uncertain, and imprecise in many ways. A major goal of the project is to map manuscript migrations, i.e., to illustrate and study manuscripts in spatio-temporal spaces using maps and and timelines, but references to locations in many cases are missing, the mentions refer to historical places that may not exist on modern maps or may have changed over hundreds of years of history [1], and initially the placenames mentioned were not even geocoded.

Also the datasets turned out to be fundamentally different in nature. The data models used in the datasets were different in different collections from TEI to relational models and RDF. But most importantly, there are substantial differences in the semantic con-

tents of the datasets: the Schoenberg Database records primarily provenance events and observations of manuscripts at specific points in time, based on, e.g., auction catalogs, and does not focus on manuscripts as unique objects, while in Bibale, Medium, and the Bodleian catalogs the main focus is on describing manuscripts as objects.

The project started by creating a list of Digital Humanities research questions relating to manuscript histories, and continued by trying to figure out what kind of data model and data are needed to solve them. The next step was to find out, given the constraints imposed by the actual data available, what questions can be addressed and under what assumptions on data. Section 3 illustrated the first steps towards this ultimate goal of the project.

Acknowledgements Thanks to Kevin Page, David Lewis and Athanasios Velios for collaborations in developing the unified data model and working on the transformations related to the Bodleian library data. Benjamin Heller developed the transformation from the Schoenberg Database format to raw RDF from which it was transformed into the unified model. Similarly, Guillaume Porte was in charge of the transformation from the Bibale database to raw RDF. Discussions with Pip Willcox, Mitch Fraas, Doug Emery, Emma Cawlfield, Antoine Brix, Synnøve Myking and other members of the project team are acknowledged.

Our work is funded by the Trans-Atlantic Platform under its Digging into Data Challenge[31] for 2017–2019. The project is led by the University of Oxford, in partnership with the University of Pennsylvania, Aalto University and Helsinki Centre for Digital Humanities (HELDIG) at the University of Helsinki, and the Institut de recherche et d'histoire des textes (IRHT). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

1. Berman, M.L., Mostern, R., Southall, H. (eds.): Placing names. Enriching and integrating gazetteers. Indiana University Press (2016)
2. Burrows, T., Hyvönen, E., Ransom, L., Wijsman, H.: 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(1), 249–252 (2018), https://mss.pennpress.org/home/
3. Clemens, R., Graham, T.: Introduction to Manuscript Studies. Cornell University Press, Ithaca (2007)
4. Doerr, M.: The CIDOC CRM – an ontological approach to semantic interoperability of metadata. AI Magazine 24(3), 75–92 (2003)
5. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/
6. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA, USA (2012)
7. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In: Proceedings of ESWC 2014 Demo and Poster Papers. Springer-Verlag (2014)

[31] https://diggingintodata.org

8. Le Bœuf, P.: Modeling rare and unique documents: Using FRBROO/CIDOC CRM. Journal of Archival Organization 10(2), 96–106 (2012), https://doi.org/10.1080/15332748.2012.709164
9. Riva, P., Doerr, M., Žumer, M.: FRBRoo: Enabling a common view of information from memory institutions. International Cataloguing and Bibliographic Control 38(2), 30–34 (2009)
10. Tunkelang, D.: Faceted search. Synthesis lectures on information concepts, retrieval, and services 1(1), 1–80 (2009)
11. Wijsman, H.: The Bibale Database at the IRHT: A Digital Tool for Researching Manuscript Provenance. Manuscript Studies. A Journal of the Schoenberg Institute for Manuscript Studies 1(2), 328–341 (2017), https://repository.upenn.edu/mss_sims/vol1/iss2/10