RESEARCH ARTICLE

JASIST WILEY

Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data

Mikko Koho[1,2] | **Toby Burrows**[3] | **Eero Hyvönen**[1,2] | **Esko Ikkala**[1] |
Kevin Page[3] | **Lynn Ransom**[4] | **Jouni Tuominen**[1,2] | **Doug Emery**[4] |
Mitch Fraas[4] | **Benjamin Heller**[4] | **David Lewis**[3] | **Andrew Morrison**[5] |
Guillaume Porte[6] | **Emma Thomson**[4] | **Athanasios Velios**[3,7] |
Hanno Wijsman[6]

[1]Semantic Computing Research Group (SeCo), Aalto University, Espoo, Finland

[2]HELDIG—Helsinki Centre for Digital Humanities, University of Helsinki, Helsinki, Finland

[3]Oxford e-Research Centre, University of Oxford, Oxford, UK

[4]Schoenberg Institute for Manuscript Studies, University of Pennsylvania, Philadelphia, Pennsylvania

[5]Bodleian Libraries, University of Oxford, Oxford, UK

[6]Institut de recherche et d'histoire des textes, Aubervilliers, France

[7]Ligatus, University of the Arts London, London, UK

Correspondence
Mikko Koho, Department of Digital Humanities, Faculty of Arts, University of Helsinki, PL 4, 00014 Helsinki, Finland.
Email: mikko.koho@helsinki.fi

Abstract

Manuscripts are a crucial form of evidence for research into all aspects of premodern European history and culture, and there are numerous databases devoted to describing them in detail. This descriptive information, however, is typically available only in separate data silos based on incompatible data models and user interfaces. As a result, it has been difficult to study manuscripts comprehensively across these various platforms. To address this challenge, a team of manuscript scholars and computer scientists worked to create "Mapping Manuscript Migrations" (MMM), a semantic portal, and a Linked Open Data service. MMM stands as a successful proof of concept for integrating distinct manuscript datasets into a shared platform for research and discovery with the potential for future expansion. This paper will discuss the major products of the MMM project: a unified data model, a repeatable data transformation pipeline, a Linked Open Data knowledge graph, and a Semantic Web portal. It will also examine the crucial importance of an iterative process of multidisciplinary collaboration embedded throughout the project, enabling humanities researchers to shape the development of a digital platform and tools, while also enabling the same researchers to ask more sophisticated and comprehensive research questions of the aggregated data.

1 | INTRODUCTION

The study of premodern manuscripts, or manuscripts produced before the age of print, is an important research area for digital humanities in medieval studies (Da Rold & Maniaci, 2015).[1]

As direct witnesses to their times and places of production, these manuscripts are a rich and complex source of critical evidence for research in a wide range of disciplines, including textual and literary studies, historical studies, cultural heritage, and the fine arts. Although each manuscript is by definition a unique object reflecting unique instances

and conditions of production, manuscripts are nevertheless semantically interrelated, containing data about the same or related textual works, authors, places and dates of production, as well as the people and institutions who created, used, and collected them over the centuries.

There are numerous online catalogs which list and describe manuscripts, as well as many digital image collections. But there is a comparative lack of interoperable infrastructure for digital manuscript metadata which can support detailed and complex research into manuscript history and provenance. The evidence base remains fragmented and scattered across data sources (Burrows, 2018). This fragmentation has made it difficult to study manuscripts across these platforms to achieve a more comprehensive, and potentially global, understanding of the role that these unique objects have played in our shared cultural and intellectual heritage.

To address these challenges, a team of manuscript scholars and computer scientists worked to create "Mapping Manuscript Migrations" (MMM), a semantic portal, and a Linked Open Data (LOD) (Heath & Bizer, 2011) service. In 2017, the team received a Round Four Trans-Atlantic Platform Digging into Data Challenge grant to combine data from several disparate sources for premodern manuscripts, and to use the aggregated data to explore a range of research questions about manuscript history and provenance.

MMM currently includes data from three harmonized heterogeneous databases of premodern manuscript metadata. Each database approaches manuscript description and metadata in markedly different ways. Bibale, a database produced by the Institut de recherche et d'histoire des textes (IRHT) in Paris, documents the history of transmission of manuscripts and their textual components from one collection to another across time (Wijsman, 2017). The Schoenberg Database of Manuscripts (SDBM)[2] (Ransom et al., 2018) at the University of Pennsylvania Libraries collects data from over 14,000 auction and sale catalogs, institutional catalogs, and other published or non-published sources, including users' personal observations. Data accuracy is dependent on these sources and may be highly variable. Although the SDBM captures detailed descriptive data about manuscripts, it is primarily used by researchers to track the ownership history of individual manuscripts.

These two databases are built on bespoke and complex data models, and each delivers highly structured data dependent on external authorities for names and places. The third source, Medieval Manuscripts in Oxford Libraries, consists of XML documents encoded according to the Text Encoding Initiative (TEI) P5 guidelines for manuscript description.[3] These XML files are a digital transcription of historical manuscript catalogs, which have been annotated with TEI tags to mark up structured data elements.

MMM stands as a successful proof of concept for harmonizing and aggregating three distinct datasets into a shared platform for research and discovery, with the potential to add more. The data harmonization uses Semantic Web (SW) technologies,[4] which have previously been successfully applied in harmonizing and publishing disparate heterogeneous cultural heritage data (Meroño-Peñuela et al., 2015). These technologies enable using shared data models and ontologies for representing distributed heterogeneous metadata from different collections in an interoperable way.

Key to the success of MMM was the back-and-forth process of collaboration among humanists and computer scientists throughout the project, from the initial identification of the research questions through the data modeling, transformation work, and interface design required to build the semantic portal and LOD service. This collaboration occurred primarily in two working groups: one focused on the requirements of manuscript provenance scholarship, and the other on the information science research needed to meet those requirements. Both groups were interdisciplinary in their membership. The first group consisted of manuscript researchers, librarians working with manuscript collections, manuscript database managers, and digital humanities specialists. The second group contained SW expertise, digital humanities expertise, and manuscript metadata expertise. Overlapping membership helped to ensure regular communication between the two groups, which contributed more than 550 person-hours' work over almost 2 years. This iterative approach ensured that major project tasks (identifying research questions, developing and implementing the data model, and publishing the data) were the product of sustained collaboration between the wide range of specialists involved, and reflected the complexity of the data and the centrality of the modeling process.

This paper focuses on four key areas where collaboration was crucial: the identification of a set of research questions; the development of the data model, based on CIDOC Conceptual Reference Model (CRM)[5] (Doerr, 2003) and FRBRoo[6] (Riva et al., 2009); a repeatable data transformation pipeline for aggregating and aligning distributed metadata into a global knowledge graph (KG); and the publishing and dissemination of the data through a SW portal and a LOD service.

The paper concludes with a discussion of evaluations carried out and lessons learned, particularly in regard to the evolving roles of the humanists and computer scientists. Though the project began with the traditional division between these two poles of digital humanities research, the outcome was a truly unified collaboration where both groups learned and benefited from the input of the other and ultimately merged into a single team, leading to a richer understanding of the harmonized data and the research questions being applied to it. Lessons

learned from this collaboration will inform the practice of manuscript description in a LOD environment to enable deeper and more integrated research into the history, and histories, of premodern manuscripts.

The paper builds on a previous dataset description article (Burrows, Emery, et al., 2020) and a paper disseminating the external vocabularies (Burrows, Brix, et al., 2020). First results of building the semantic portal have been presented in (Hyvönen, Ikkala, et al., 2019).

2 | DEVELOPMENT OF THE RESEARCH QUESTIONS

From the beginning, the goal of the MMM project was to create an online environment that would allow manuscript researchers to query data from multiple platforms. The challenge in building such an environment revolves around three key problems:

1. How to model manuscript metadata for interoperability, with a focus on provenance data?
2. How to extract, integrate, and reconcile heterogeneous manuscript metadata from distributed data sources?
3. How to publish Linked Data for digital manuscript research?

Before these questions could be answered, the project team needed to come to a shared understanding of the content and nature of the data derived from the three sources. This extended beyond the contemporary formats of the constituent datasets from a purely technical perspective, to encompass an appreciation of their heritage—the sequence of historical motivations, actions, and constraints which led the catalogs to their current configurations (in some cases, going back over 150 years). Having an understanding of the significance of the nuances and intricacies as well as the inconsistencies and ambiguities of the three datasets was essential for constructing a unified data model that was comprehensible to and useful for manuscript researchers. This was especially important given a requirement for the harmonized data service to supplement, rather than replace, the three data sources: each remains the canonical catalog for the records held within, and for their updating, with the Linked Data Service taking a layered approach (Page et al., 2017) to provide additional discovery, analysis and visualization functions.

The first stage of the collaboration therefore required manuscript scholars and metadata specialists to introduce the data modelers to the data and to identify the key data elements that would be most useful to intended users. A set of 25 sample research questions was developed to guide this work; the full list is given in (Burrows, Pinto, et al., 2020). The research questions were developed by manuscript researchers, curators, and librarians involved in the project team, based on (a) their own interests and areas of expertise, (b) the discussions of the Oxford focus group held at the start of the project, and (c) a similar but more generic set of questions developed by the Biblissima project (Frunzeanu et al., 2016). Semantically, they ranged from relatively straightforward ("How many manuscripts were produced in London in the fifteenth century?") to considerably more complex ("How many surviving manuscripts that contain Spanish texts written in gothic rotunda were produced in Castile for an abbey or convent?").

The source datasets were then analyzed together with the research questions to identify key data points in manuscript description that would provide the building blocks for a unified data model. This process identified the following elements: author, title, language, place of production, date of production, script, provenance agent, provenance event time range, and current location. Once these desired elements were established, the next step was to confirm that they existed across all three datasets and could be programmatically extracted. If all three datasets contained the same elements in a research question, then it was likely to make a viable contribution to the unified model. If only one or two of the datasets contained an element, then the potential viability of the question for the unified model needed to be reconsidered. For example, a common but not consistently used factor in manuscript description is the identification of the script, or style of writing used by a scribe, such as Carolingian miniscule or Gothic textura. Neither the SDBM or Bibale captures this kind of data, so this element was omitted from the model.

Analysis of the research questions also exposed semantic flaws in the questions that would impact the data model development. Several questions contained a significant level of ambiguity. For example, the question "What was the most popular text by a medieval author in France in the seventeenth century?" poses computational problems because the concept of "popular" is ambiguous. This question, like several others, also challenges modern assumptions about concepts like "text" or "author." Many medieval texts do not have titles or authors in the modern sense, and the attribution of title and author to a given text may also have changed over the centuries (Sharpe, 2003). Compilations, translations, redactions, and other editorial permutations further complicate the notion of authorship and the concept of what constitutes a "work" as defined in FRBRoo, for example.

An awareness of ambiguities and inconsistencies in the data at this point in the process flagged potential semantic problems that the data modelers could then take into account, which will be discussed in the MMM Data Model subsection. The results of the review showed

where the data elements aligned well and where they did not, as well as revealing some notable gaps in expected elements of manuscript description apparent in all three data sets such as script. Once the elements were identified, they were then mapped to the research questions. This mapping identified those elements that were most critical for answering the questions or that would support the most robust querying.

Within our iterative collaboration, the research questions first provided the structured requirements from which to begin modeling the data, and thereby the scope which determined those elements which would be necessary to include within the unified model (and, conversely, those elements which were out of scope for the requirements of this project, and could be left for future extensions). The research questions continued to serve as a common foundation for inter-mediation between the expectations of manuscript scholars and the affordances of the data model as it evolved, providing the basis for a reciprocal process of development. The research questions enabled development of the data model, while the developing data model also fed back into the refinement of the research questions. Across the project there emerged a mutual respect for the data modeling process as an iterative and continuous intellectual collaboration between content expertise and technical expertise.

The set of research questions was also integral to testing the data harmonized with the unified data model. The "scholarship group" employed the questions to explore the user interface and provide feedback to the developer and designer. In several cases, the feedback also led to revisions to, and extensions of, the data model itself (e.g., in such problematic areas as "last known location" of a manuscript). From July 2019, a group of researchers, librarians, manuscript metadata specialists, and computer scientists held weekly workshops in which they used the MMM SPARQL endpoint to query the data directly, once again taking the set of research questions as a starting-point for queries. These explorations also contributed to further refining of the data model.

3 | DATA MODELING

Data from the three source databases (Bibale, SDBM, and Medieval Manuscripts in Oxford Libraries) was harmonized and published as LOD. For this purpose, a harmonizing data model was devised, based on CRM and FRBRoo. CRM is an event-based model for information integration in the field of cultural heritage. FRBRoo is a CRM compatible version of *FRBR*, a conceptual model for bibliographical information (Le Bœuf, 2012). Building on earlier, related work, the MMM project nevertheless developed a new and unique data model which combines elements of CRM and FRBRoo with MMM-specific elements. Sub-properties of some CRM and FRBRoo properties are used for more specific relations, while also completely new properties are added for data that is out of the scope of CRM and FRBRoo, for example, *mmms: last_known_location* which is used to refer to the last known location of a manuscript, based on the available data.

3.1 | Related work

Modeling rare and unique documents like manuscripts using CRM and FRBRoo has been studied in Le Bœuf (2012), the insights of which have guided our data modeling work. The suitability of CRM for representing manuscript metadata has also been noted in Bellotto (2020). There are some existing SW approaches for harmonizing manuscript collections using CRM and FRBRoo, including the Biblissima project (Frunzeanu et al., 2016; Gehrke et al., 2015) and a catalog of historical Hebrew manuscripts (Zhitomirsky-Geffet et al., 2020), The Biblissima project has developed a data model based on CRM and FRBRoo for the purpose of integrating and harmonizing a number of heterogeneous manuscript databases as LD. Although their data model is much broader in scope, the data model and data mapping templates were used as inspiration for the MMM data model.

Additionally, the Europeana Data Model has been extended[7] for integrating manuscript metadata collections as LD into the Europeana data portal (Baierer et al., 2017), providing an alternative basis for manuscript metadata harmonization.

Another model for medieval manuscripts is the Medieval Manuscripts Ontology (MeMO) (Barzaghi et al., 2020), which is designed for modeling the metadata of the digitized medieval texts of the Royal College of Spain in Bologna. MeMO was based on FRBR, but CRM was disregarded as "over-engineered and too difficult to comprehend" for this specific use case. This choice illustrates the fundamental difference of flexibility in developing a data model only for a single catalog, instead of having to adapt to multiple heterogeneous catalogs.

3.2 | Data sources

As noted in the Introduction, the MMM KG consists of data from three databases: Bibale, the Schoenberg Database of Manuscripts (SDBM), and Medieval Manuscripts in Oxford Libraries. While all three focus on premodern manuscripts, each database serves a different purpose

and each has followed its own approach to manuscript description.

3.2.1 | Bibale

Bibale's rich data model supports the representation of detailed information about individual objects and their associations. All Bibale records belong to one of eight object types: manuscripts, works, persons, bindings, collections, ownership marks, texts, and sources. A record can be associated indefinitely with any other record (including one belonging to the same object), e.g., a person with another person to say that they are father and son; a manuscript with another manuscript to say that they were once bound together or have been copied one from the other; a work to another work to say that one is a translation of the other; and so on.

Bibale contains roughly 55,000 records (December 2020) representing the following data:

- 13,750 persons (more than 2,000 of which are institutions).
- 3,750 collections (private or public libraries).
- 17,000 books (almost all manuscripts, some printed books).
- 750 bindings.
- 13,500 provenance marks (ex-libris, ex-dono, heraldic arms, etc.).
- 500 external sources (ancient catalogs, inventories, lists, etc.).
- 1,250 texts or editions (the version of a text found in a specific manuscript).
- 2000 works.

Bibale contains references to external authorities, including VIAF for name authorities and textual place references based on GeoNames. The database contents were exported as CRM for the MMM project.

3.2.2 | Medieval Manuscripts in Oxford Libraries

As of January 2021, the Medieval Manuscripts in Oxford Libraries dataset covers 10,272 manuscripts. Most of the descriptions are summary entries, digitally encoded from the Quarto and Summary Catalogs published between 1853 and 1924. Detailed modern descriptions are available for manuscripts acquired since 1916. The manuscript descriptions and authority files are encoded in XML according to a customization[8] of the Guidelines of the Text Encoding Initiative (TEI). Significant effort has been invested in the creation of local authority files for works, people and places, also using TEI. These have been, in turn, manually reconciled with uniform resource identifiers (URIs) of records in external authorities such as VIAF, Library of Congress, Bibliothèque nationale de France, Système Universitaire de Documentation, Gemeinsame Normdatei, and WikiData. While the XML gives structure to the manuscript descriptions, some of the data that would become important to MMM, in particular the provenance data, is contained in narrative note fields; this could not be easily extracted and would require the team to find a workaround.

The XML files, together with processing software, are stored in a publicly accessible GitHub repository.[9] The Bodleian's own catalog website[10] is generated from a version controlled check out of the XML repository. For the MMM project, the XML files were transformed into Resource Description Framework (RDF) based Linked Data (Burrows et al., 2021). The first step was to extract a selection of the XML elements from each file and combine them into a simpler, flatter XML structure using xQuery. A transform from the simplified XML into CRM compatible RDF was generated in the 3M (Mapping Memory Manager) software,[11] then scripted as a reproducible Docker file which upload the resultant data to a git repository.

3.2.3 | Schoenberg database of manuscripts

Entries in the SDBM use 36 possible fields to record data from observations of manuscripts found in published and unpublished sources. The database contains over 250,000 records containing provenance-related observations of manuscript histories. The history of SDBM traces back to Lawrence J. Schoenberg, who started building the database in 1997 for private use. Currently the SDBM data is stored in a MySQL relational database.

The current data model, launched in 2017,[12] is focused on the entries (manuscript observations) and their sources. Sources, such as auction catalogs and bookseller websites, describe manuscripts. Entries represent observations of manuscripts derived from a Source. A Manuscript Record links together entries that describe the same manuscript, gathering observations across time and place for easy reference and study. SDBM uses Virtual International Authority File (VIAF) based Name and Place Authority Files to standardize spelling and naming conventions for people, organizations, and places. The place records are also linked to Getty Thesaurus of Geographic Names (TGN) and GeoNames identifiers. The SDBM also makes an RDF version[13] of the dataset available via a SPARQL endpoint,[14] which supplies the initial conversion for the MMM project.

The SDBM data model presented two problems for MMM. First, multiple entries can represent a single manuscript. An entry typically corresponds to an auction lot or catalog entry. The same manuscript can be represented by multiple entries because manuscripts move through time and can be "observed" in different sale or collection catalogs. Many entries in the SDBM are linked together to form Manuscript records, but linkage among entries is not comprehensive. It is therefore unclear whether unlinked entries represent a unique manuscript or have simply not been linked to a manuscript record. A second challenge relates to the concept of intellectual works contained in the manuscripts. In the SDBM data model, Authors are not conceptually linked to Titles but treated as unrelated data elements. This discrepancy would eventually require the MMM project to perform a reconciliation process to match titles and authors manually.

In spite of the different approaches to manuscript description and related metadata described above, each data source bore enough structural overlap to make unification possible, thanks to links to external authorities for Names and Places, as well content overlap in key data elements. These areas of overlap, which boiled down to Manuscripts, Works, Actors, Places, and Events, also eventually shaped the interface design.

3.3 | Modeling manuscript metadata as linked data

Combining manuscript provenance metadata in an interoperable way from the heterogeneous data sources is required to depict a more complete view of the histories of the manuscripts. The semantic reconciliation needed for this task requires making the data semantically interoperable (Hyvönen, Ahnert, et al., 2019). This requires both the reconciliation of schemas and the reconciliation of entities in the data sources.

Schemas can be reconciled by means of devising a unified data model encompassing the relevant entities and relations in the source datasets, to overcome their own data modeling conventions, and mapping the datasets into this schema. The MMM data model is based on CRM and FRBRoo, which support event-based modeling needed for provenance data that are essentially chains of events concerning the manuscript objects.

The FRBR model makes essential distinctions between four distinct layers: an abstract intellectual *Work* can be available in various *Expressions* (e.g., different languages), contained in physical *items* that belong to the same *manifestation* (Le Bœuf, 2012). FRBRoo is more suited to modeling unique documents, such as

manuscripts, than the original FRBR model, by allowing the accounting of the histories of the documents through various CRM-based events (Le Bœuf, 2012). The FRBRoo class Manifestation Singleton corresponds to the notion of unique documents, whereas the Item class is only used for documents produced in multiple copies (Le Bœuf, 2012). A manuscript carries its intellectual content in a Self-Contained Expression, or a fragment of a Self-Contained Expression, which in turn realizes a Work (Le Bœuf, 2012).

The production of a manuscript is expressed in FRBRoo differently depending on whether it is the production of an original manuscript, or the production of a copy of another manuscript (Le Bœuf, 2012). The activity F28 Expression Creation corresponds to the former, resulting in new instances of each F2 Expression and F4 Manifestation Singleton. The production of a copy of a manuscript can also produce some difference from the original, and thus could be considered creating a new expression. However, the data available and the objectives in studying the data should dictate whether it makes sense to model the differences in expressions or consider copies to be exact copies (Le Bœuf, 2012).

In harmonizing the information in the source datasets, reinterpreting the source information in terms of the unifying data model is necessary. In some cases, we are required to make assumptions about the contained entities as there is not enough information to follow the FRBRoo model fully. The strict separation of information about the intellectual contents of a manuscript into Work and Expression levels can be difficult in practice. These levels are not modeled separately in any of the data sources and this distinction is difficult to impose in the reinterpretation phase. Some information generally exists on both levels, underpinning the importance of this challenge.

3.4 | The MMM data model

The MMM harmonizing data model uses the Erlangen versions[15] of CRM and FRBRoo. An overview of the data model is presented in Figure 1, in which the nodes represent classes and the arrows represent commonly used properties between the instances of those classes. The MMM data model makes also use of subclasses and subproperties of those shown in the figure. The namespace prefixes used in this article are shown in Table 1.

The scope of the MMM data model is restricted to the information available in the data sources, but the data model can easily be extended as needed, if new datasets are to be harmonized with the MMM data.

Following the good practices of using CRM in RDF (Doerr et al., 2020), we do not model names as
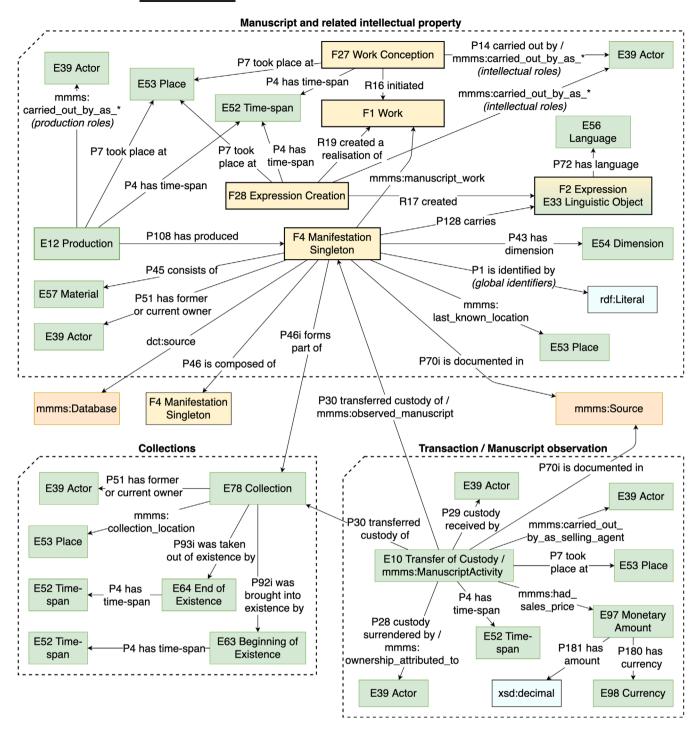
FIGURE 1 The MMM data model for harmonizing manuscript metadata. The core classes of the MMM data model shown as rectangles and common properties between class instances shown as arrows. The FRBRoo classes are yellow and CRM classes are green. The properties starting with a letter and number refer to FRBRoo and CRM properties. The instances of the F4 Manifestation Singleton class correspond to individual manuscripts

appellations, but rather use literal values directly as names using SKOS properties. Similarly, the CRM time primitives are not used, but instead temporal extents of time-spans are quantified by date references using four properties (P82a, P82b, P81a, P81b).

Sources for individual pieces of information are given on the resource level as *dct:source*, indicating the source dataset of the resource. A URL link to the resource in the user interfaces of the source datasets is also added to the data where possible.

TABLE 1 Used namespaces prefixes

Prefix	URI	Name
crm	http://erlangen-crm.org/current/	Erlangen CRM
dct	http://purl.org/dc/terms/	DCMI Metadata Terms
frbroo	http://erlangen-crm.org/efrbroo/	Erlangen FRBRoo
mmms	http://ldf.fi/schema/mmm/	MMM Schema

3.4.1 | Manuscripts

Manuscripts are modeled as instances of *frbroo:F4_Manifestation_Singleton*. As there is no information available whether individual manuscripts are copies or original pieces of work, manuscripts are always assumed to be copies of other manuscripts and hence produced by a *crm:E12_Production* event.

The databases contain three different notions of ownership of a manuscript. (a) There is ownership that can be observed from visual signs in the manuscript itself, for example, ex-libris and seals. (b) Additionally, there is ownership that can be observed from a collection catalog. (c) In auction catalogs and some other sources, there can be information about an event in which a manuscript has been obtained or given away by someone. These are all modeled differently in the data model, with (a) expressed as a direct ownership of the manuscript resource, (b) expressed via a Collection resource, and (c) expressed as 1, with the addition of a Transfer of Custody event.

The concept of a manuscript as a physical unit is inconsistently used, often even within catalogs. A manuscript in a database can be one of several different manuscript concepts:

1. Manuscript group: several physical things that are grouped together, probably because they are volumes of a single manifestation of a work.
2. Manuscript volume: a single physical object—often what people, especially in libraries, mean by manuscript.
3. Manuscript part: a constituent part of a manuscript volume which has at some point been physically separated from the other parts. If a manuscript volume has a part in it, then all of the volume should be described as parts.
4. Manuscript fragment: a physical thing that was at some point a constituent part of a larger manuscript volume and not an independent part.

Of the source datasets, the Oxford catalog has separated these different levels in its data, whereas the SDBM and Bibale databases only address this to a limited degree. In the MMM data these are handled as part-of relations (*crm:P46_is_composed_of*) between manuscripts when it is known that a manuscript is a part or a fragment of a manuscript group or volume.

All known owners of a manuscript are expressed with the property *crm:P51_has_former_or_current_owner*. In addition to this, more detailed provenance information is often available through events related to the manuscript.

3.4.2 | Works and expressions

Works are modeled as instances of *frbroo:F1_Work* and their textual expressions as instances of both *frbroo: F2_Expression* and *crm:E33_Linguistic_Object*.

In the SDBM data, the authors of manuscripts and the titles contained in the manuscript are known. However, these are not linked to each other. This vague connection between the authors and titles is problematic to express in FRBRoo terms. Each title of a manuscript represents both a separate expression and a work, as there is some data of both levels, but the authors (and scribes) are expressed for the Work Conception and Expression Creation events with a custom property *mmms:carried_out_by_as_possible_author*, which means that the actor may have been involved in it. The helper property *mmms:manuscript_work* connects a manuscript to all of its known works and the helper property *mmms:manuscript_author* connects a manuscript to all of its known authors.

3.4.3 | Events

Many details about manuscripts and their provenance are modeled as events, which are all subclasses of *crm:E5_Event*. Places and time-spans are expressed always when they are known, with *crm:P7_took_place_at* and *crm:P4_has_time-span*, respectively. Provenance information is expressed in *crm:E10_Transfer_of_Custody* (change of manuscript ownership) and *mmms:ManuscriptActivity* (observations of manuscripts) events.

Actors taking part in an event are given with various CRM based properties. There are eight different actor roles (author, artist, scribe, binder, etc.) in production, expression creation, and work conception events, which are modeled with role-specific sub-properties of *crm:P14_carried_out_by* such as *mmms:carried_out_by_as_author* and *mmms: carried_out_by_as_scribe*.

3.4.4 | Actors

Actors are modeled as instances of *crm:E39_Actor* or its subclasses *crm:E21_Person* and *crm:E74_Group*. Main

information known about the actors are their various labels as *skos:prefLabel* and *skos:altLabel* and the birth and death dates of persons. Additional information present in some cases include places of birth and death, place associations via nationality or residence, or gender as *mmms:gender*. Most of the information is accessible through various related events.

3.4.5 | Places

Geographical information has been expressed in each source dataset according to their individual conventions. The SDBM hosts a well-structured place authority[16] where external vocabularies are prioritized in the following order: (a) Getty TGN, (b) GeoNames, (c) VIAF. The Oxford catalog prioritizes Getty TGN, but additionally makes references to GeoNames and Historical Gazetteer of England. When populating Bibale database the place references (i.e., Country–Region–Settlement) were taken from GeoNames, but the GeoNames ID was not stored in the database.

Based on this initial state, it was decided that the Getty Thesaurus of Geographic Names[17] (TGN) would serve as the best shared place authority for all three databases. TGN was chosen because it was already used in most source datasets, and it is to our knowledge the only widely adopted digital gazetteer with a global coverage and support for the temporal dimension of geographic information.

As the CRM and FRBRoo models do not offer any built-in models for geographical information, the primary data model for geographical information in MMM is the Linked Data model[18] of TGN.

4 | DATA HARMONIZATION AND TRANSFORMATION

The data model is populated by reinterpreting the source datasets in terms of the data model. A repeatable automated data transformation pipeline[19] was developed to (re-)create the whole MMM KG from the source datasets when needed. To facilitate reproducibility, the pipeline is based on Docker. The pipeline enables updating the KG regularly with updated source datasets.

The functionality of the pipeline is depicted in Figure 2. The three data sources in their original formats are depicted in red. The pipeline takes the RDF exports of the source datasets as input and first transforms these into the MMM data model using a total of 22 SPARQL Construct queries. The queries reinterpret the information in the source datasets in the terms of the MMM data model.
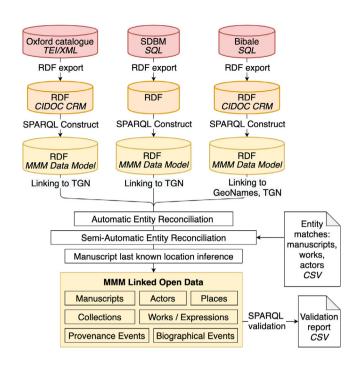

FIGURE 2 The MMM data harmonization and integration pipeline

After the databases are transformed into using the MMM data model, various processes are used to disambiguate and reconcile a portion of the entities shared between them. Entity reconciliation processes merge entities both automatically and based on CSV files of matches created by domain experts. The matching entities are reconciled by merging the entities and their metadata, while also pointing all references to the new entity.

To support data visualization, last known locations of manuscripts are inferred from the data and annotated to each manuscript. After the transformation and reconciliation tasks are completed, a set of 3 SPARQL queries are used to validate the data to find possible errors.

4.1 | Automatic entity reconciliation

This step combines automated processes for matching and reconciling entities originating from different source datasets. This step aims to reconcile all places between the source datasets, but provides non-exhaustive results to the reconciliation of manuscripts and actors.

The first step of this process was to identify the vocabularies used in the source datasets and how far the reconciliation is possible using these vocabularies. This revealed some gaps in the use of vocabularies, which were addressed to facilitate the automatic reconciliation process. Finally, automatic reconciliation processes were developed, which use the shared vocabularies.

Duplicate actors from multiple data sources are recognized automatically based on their shared VIAF identifiers. Automatic processes match 589 actors between the databases based on their VIAF identifiers.

Collection shelf-mark identifiers assigned to manuscripts and quoted in database descriptions can be used to automatically reconcile references to the same manuscript. Automatic linking can also be carried out based on the Phillipps numbers assigned to manuscripts formerly belonging to the huge collection of Sir Thomas Phillipps (1792–1872) (Munby, 1960). There are a number of problems with this approach. A single Phillipps manuscript volume can be divided into multiple manuscript volumes in the datasets. For example, the manuscript with Phillipps no. 10584 corresponds to 38 items in the Bibale database. Conversely, there exists a single manuscript in Bibale that used to consist of two separate items in the Phillipps collection, and hence has two different Phillipps numbers. 3,525 matches between manuscripts are found by matching based on shelf-mark identifiers parsed from the manuscript metadata (3170) and Phillipps numbers (355).

The textual place references of GeoNames places in Bibale are linked back to GeoNames. One issue with this is the temporal changes that have occurred in GeoNames since the textual references have been made, requiring to take into account the French administrative region reform in 2016. After linking to GeoNames, the places are linked to TGN. Then all place references from all datasets are harmonized by creating shared resources for the TGN places.

4.2 | Semiautomatic entity reconciliation

Semiautomatic entity reconciliation was done employing the domain knowledge contained within the project. This was based on the *Recon*[20] tool, which is designed for digital humanities scenarios where trusted accuracy is important (Hyvönen, Ahnert, et al., 2019) and hence entity matching cannot be done completely automatically, but instead a person is consulted to consider possible candidates for matching. Initially possibly useful reconciliation scenarios were devised, based on the domain expertise and computational possibilities for finding possible candidates for matching. For each scenario, a predefined configuration and SPARQL queries were created to provide the domain experts with candidates to reconcile manually.

In the matching, entities were considered unique in a single source dataset and matched only between the datasets. The match candidates were scored and sorted

and relevant metadata about the entities, including links to the data in the source datasets, was shown to the user to make an informed decision. In this fashion, the project matched 3,136 actors and 1,067 works between the source datasets.

Additionally, a number of manuscripts were reconciled manually by manuscript scholars in a spreadsheet, instead of using Recon. Eighty manuscripts were matched between two or three of the databases in this way.

5 | PUBLISHING THE HARMONIZED KG

When dealing with scholarly data, the MMM system follows the "FAIR guiding principles for scientific data management and stewardship"[21] To enable data-driven manuscript research and facilitate data reuse, the harmonized dataset of 24 million RDF triples is available in several ways:

1. In the Zenodo repository[22] with a canonical citation (Koho et al., 2021);
2. Hosted on the LDF.fi platform,[23] from where it can be accessed via a SPARQL endpoint or using a web browser;
3. On the online MMM Portal,[24] through which the 222,600 manuscripts and other entities can be searched and browsed.

5.1 | Data service

The MMM KG is available on the Linked Data Finland (LDF) platform (Hyvönen et al., 2014), providing a home page for the KG, and a public SPARQL endpoint.[25] To support reuse, the home page provides additional information about the KG, such as, (a) schema documentation automatically generated by the platform, (b) sample SPARQL queries, and (c) metadata using SPARQL Service Description,[26] and Vocabulary of Interlinked Datasets (VoID).[27]

The MMM SPARQL endpoint is hosted on an Apache Jena Fuseki[28] SPARQL server. The whole KG and Fuseki are contained in a Docker image, that can be easily built and started when and where needed.

The LDF platform provides dereferencing of URIs for both human users and machines, and a generic RDF browser for technical users, which opens when a URI is visited directly with a web browser.

5.2 | MMM portal

The goal of the MMM Portal is to enable large-scale exploration of data relating to the history and provenance
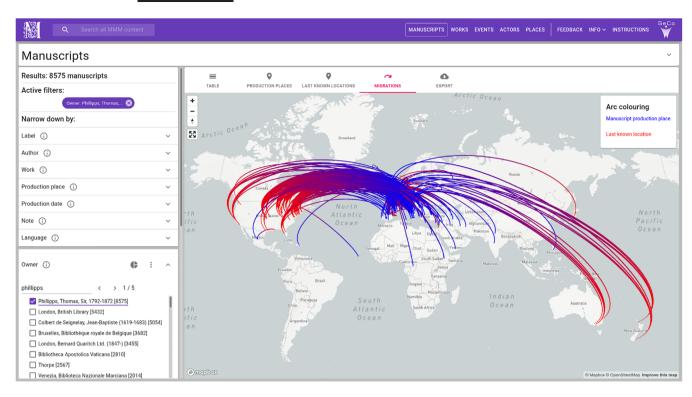
FIGURE 3 Visualization of the movement of premodern manuscripts from places of production to last known locations using the MMM Portal

of premodern manuscripts. The main user group is manuscript researchers, who need to be able to analyze and visualize the harmonized and aggregated data at scales ranging from individual manuscripts to hundreds of thousands of manuscripts or other entities of interest.

The portal provides five perspectives to the MMM KG, based on the main entity classes: Manuscripts, Works, Events, Actors, and Places. The perspectives are equipped with faceted search (Tunkelang, 2009) and browsing engines integrated with ready-to-use visualization tools for Digital Humanities research.

After choosing a perspective, the user is presented with an initial result set, which contains all entities of the class at hand. By default, the result set is displayed as a paginated table. Starting from the initial result set, the user can narrow it down using a set of predefined facets. For example, the hierarchical Production place facet in the Manuscripts perspective can be used to narrow the initial result set to manuscripts produced in Europe, based on the hierarchy provided by the Getty TGN. In addition to faceted search, the user can also perform a full text search for all entities in the MMM KG, using the search field which is always visible on the top navigation bar.

It is also possible to create alternative analytic visualizations for the result set, too, represented as separate tabs. For example, in the Manuscripts perspective there are separate tabs for visualizing the production places, last known locations, or migrations of manuscripts as interactive maps. The visualizations are automatically updated when the user changes facet selections. Figure 3 illustrates how faceted search and the migrations tab can be used to study the movement of manuscripts from their place of production to their last known location. Using the Owner facet on the left-hand side, the initial result set of 222,605 manuscripts has been narrowed down to manuscripts that once were owned by Sir Thomas Phillipps. The blue end of the arc shows the production place, and the last known location is shown in red.

The harmonization work of the MMM project enables representing each entity of the main classes as a clickable link when visualizing the result sets. The links lead to entity landing pages, which list all relevant information about the entity. This simple principle interlinks the five perspectives of the MMM Portal extensively with each other, thus providing the user a flexible and user-friendly mechanism to study the underlying KG.

The user interface of the MMM Portal is implemented with the Sampo-UI framework (Ikkala et al., 2021). This means that the user interface is a web-based application written purely in JavaScript. The application consists of a NodeJS[29] backend built with Express framework[30] and a client based on React[31] and Redux.[32] The client contains all logic for displaying the data and reacting to

user's selections. In order to fetch data, the client makes use of the MMM SPARQL endpoint by sending API requests[33] to the Node.js backend. Based on the API requests and predefined configurations, the Node.js backend generates the SPARQL queries, sends them to the SPARQL endpoint, maps and merges the raw result rows, and returns the results as nested JavaScript Object Notation (JSON) structures to the client, where the data is interpreted and visualized to the user.

5.3 | Using the KG

The data analysis tools and visualizations of the MMM Portal combined with faceted search provide an easy-to-use starting point for analyzing manuscript provenance. In addition to this, one can use the public SPARQL endpoint to explore and analyze the data.

The information in the MMM KG contains several benefits compared to the source datasets:

- The user is able to execute arbitrary SPARQL queries over the whole KG, instead of being limited to using the predetermined search user interfaces of the individual source datasets;
- Harmonized basic unit for manuscripts (FRBRoo class Manifestation Singleton). For example, the user cannot search for or count manuscripts in the SDBM database, but only observations of them (e.g., entries in sales catalogs);
- Reconciled and enriched actors and places across source datasets. For example, from the Bibale database, plotting manuscripts or events on a map is not possible;
- Last known locations are not available for most of the manuscripts in the source datasets. These have been inferred based on the harmonized data.

The MMM data service and SPARQL endpoint can be used by any external application or tool. For example, (Klyne & Lewis, 2020) provides a report on exploring the MMM KG using ResearchSpace[34] for addressing specific research questions, such as "Who collects manuscripts with texts by Ramon Llull?" Data relating to more than 8,000 manuscripts formerly owned by Phillipps have been extracted via SPARQL and reused in a *nodegoat* database environment.[35]

6 | EVALUATION

An initial evaluation of the MMM aggregated data was carried out using the 25 research questions developed for the project. Each question was tested first against the three source datasets individually. In almost every case answering the questions fully proved difficult. At best, the user was presented with a partial answer to the question, often as a broader list of results which had to be scanned manually to identify relevant items. Some questions could not be answered using the source datasets (8 in Bibale, 8 in Oxford, 6 in SDBM). When the same questions were run against the MMM Portal, 17 out of 25 could be answered with a combination of filters and text searches. Only a few, more complex, questions required further manual scanning of the result sets (8 out of 25). This group of questions was then explored further by running queries against the MMM SPARQL endpoint. This evaluation demonstrated that the aggregated MMM dataset can support more sophisticated and complex queries from researchers than the source datasets. The results of this evaluation are summarized in Table 2.

One of the more complex research questions was: "What French collectors purchased manuscripts since the end of the Wars of Religion (after 1598)? Where are their manuscripts now?" This cannot be answered in Bibale or the Oxford catalog. In Bibale, it is impossible to run a query on transactions of a specific period, while in the Oxford catalog the list of people can be filtered by role (e.g., owner) but not by place. SDBM does make it possible to identify people linked to France with life dates after 1,598, and then view the individual entries linked to them. But this will not cover people linked to specific places within France, since place names are not nested hierarchically.

In the MMM Portal, on the other hand, the "Actors" perspective can be filtered for persons with an "Activity Location" of France. This covers all places within France. Finding French collectors active after 1,598 involves adding one of the timeline filters to find persons born after (say) 1,550. The resulting list of 572 people includes a list of manuscripts and collections attached to each of them. These manuscripts and collections can then be inspected to see their subsequent history and last known locations. The list of people can be sorted by "Role" to distinguish manuscript owners and collection owners from authors of works. To amalgamate all the relevant information about each manuscript and each collection for each owner who falls within the specific parameters, a SPARQL query can be constructed.

The ability to find interesting knowledge from the MMM Portal has been noted in Engels (2020). An evaluation of the MMM Portal by three postdoctoral manuscript researchers is reported in Burrows, Pinto, et al. (2020). As well as providing detailed feedback on the functionality of the Sampo-UI interface, this report made recommendations for fuller public documentation. It also identified

	Bibale	Oxford	SDBM	MMM portal
Impossible to answer	8	8	6	0
Partly answered	16	12	12	8
Fully answered	1	5	7	17

TABLE 2 Answers to MMM research questions from individual datasets and the aggregated data

some specific issues with the display of results in the portal. These issues arose largely from ambiguities and inconsistencies in the data in the source datasets, rather than errors in the mapping and transformation processes.

Among these issues identified was the occurrence of multiple production places for the same manuscript. This might result from disagreements between the data sources, but was more like to reflect uncertainty about its origins: Southern France or Northern Italy? Multiple and sometimes conflicting production date ranges were also fairly common, reflecting differing opinions among cataloguers and scholars. Some authors had multiple different entries in the list of persons, whereas others had only one entry, with all the variant versions harmonized; this was the result of incomplete coverage in the reconciliation process. Similarly, the titles of works were only rarely reconciled, making it very difficult to track specific textual traditions—although this was never one of the goals of the project.

Dealing with ambiguous and inconsistent naming conventions related to geographical regions was another challenge. For example, the term "Northern Italy" is used in all source datasets, but it is impossible to interpret automatically which specific regions and cities this term covers. There may also be multiple varying conceptions of Northern Italy, caused by varying data curation practices.

Inconsistent naming conventions and the granularity of the available data affect the analysis of geographical distributions. For some manuscripts there may be information about the exact monastery where it was written, whereas in many cases the only geographic reference available is either a city, a larger region, or even only a continent. Still, demonstrating the potential of geographical analysis and visualizations within the MMM project was a reasonable result and has also encouraged the curators of the source datasets to improve and standardize the original geographic references, thus returning a benefit of the project to the source datasets.

7 | CONCLUSIONS AND LESSONS LEARNED

This paper presented an approach for harmonizing heterogeneous manuscript metadata databases with a focus on the manuscript description and provenance, leading to four primary results. These are:

1. The design of a proof of concept for an event-based harmonizing data model capable of representing both manuscript metadata and their provenance information in detail, that can also be used for integrating heterogeneous local manuscript datasets into a global KG.
2. The design of a production model for aggregating and harmonizing distributed heterogeneous datasets into a global Linked Data Service, based on the harmonizing data model and a set of shared vocabularies for populating the model.
3. A successful example of a style of iterative and discursive collaboration among manuscript scholars, metadata specialists, digital humanists, and computer scientists that informed the development of the data modeling, the processes of transformation, and the design and defined purpose of the interface as an entry and initial guide to the exploration of the combined datasets in the unified KG.
4. A better understanding of the behavior of the three different approaches used in the source datasets to manuscript description and metadata in a SW environment and recommendations for future projects involving digital manuscript description.

The semantic portal demonstrator MMM and its underlying LOD service created and published on the SW provide evidence for the success of the first two results. In the process of developing the portal and the service, the project also produced the following artifacts that have the potential to impact future work that may build upon the MMM model. These are:

- A unifying data model for manuscript metadata and provenance;
- Data transformation pipeline and tools for linking and harmonizing the source datasets;
- A LOD publication of the integrated data;
- Vocabularies for four main entity categories reconciled across the data sources: Manuscripts, Works, Actors, Places;
- LOD vocabularies with unique identifiers for 222,605 manuscripts, 435,428 works, 5,077 places, 56,685 persons and organizations, and 1,880,399 events.

Harder to quantify but no less important are the third and fourth primary results. The fruitful collaboration between the two working groups identified at the outset—the first group focused on the requirements of manuscript provenance scholarship, and the other on the information science research needed to meet those requirements—was key to the success of the project. The ability to maintain open lines of communication throughout the development process as it progressed iteratively from identification of research questions, to modeling design and implementation, and finally to publication and evaluation, provided the mechanism to build a robust, transparent, and intuitive resource for manuscript discovery. The challenges presented by the three datasets were overcome through careful analysis of them by the "scholarship" group that was then communicated to the "information science" group to provide building blocks for designing the first iteration of the data model. Upon implementation of the model, both groups worked together to apply the initial research questions to test the model.

Testing the model through extensive SPARQL querying produced two significant results. It allowed the "information science" group to fine-tune the data model in response to particular successes or roadblocks discovered in the application of the initial research questions. Second, the "scholarship" group became more proficient in querying the data, which encouraged them to develop new research questions not previously considered. Testing thus gave the "scholarship" group greater clarity about the contents, limitations, and new possibilities for research and discovery presented by the combined dataset. For example, a query to determine the average ratio of the page size of liturgical manuscripts produced between 500 AD and 1,601 AD produced 4,498 results. The query is available[36] in the online Yasgui tool, which is also used to visualize the results. Presented in the scatter plot graph of Figure 4, the data shows an expected average height to width ratio to be between 1.5 and 2.0 for liturgical manuscripts in codex form. More interesting are the three outlier points where the ratio is greater than 8.0. Two instances between the ratios 8.0 and 11.0 are the result of data entry error in the source datasets. The highest ratio, however, points to an atypical context for a liturgical manuscript, a roll in which the text would begin at the top and continue down the face of the roll (its recto). The text is a set of prayers from a breviary, a book that outlines the liturgical regime of religious institution over the course of a year. A breviary is typically held in the hands for an individual's reference. A roll, therefore, is truly atypical for such a text and suggests a different functional context. The data both confirms that this format is atypical and begs further academic investigation into the circumstances that produced such a format.

This iterative approach to collaboration among the different specialists involved in the project is a contribution to ongoing discussions about the challenges involved in modeling data in the humanities (Zöllner-Weber & Apollon, 2008) and about ways of getting humanities researchers to work with structured data (Breitenfeld et al., 2018). In particular, the project provided an instance of the kind of approach outlined and recommended in Oldman et al., 2016: "The advanced methods of the RDF/OWL framework to express meaning and to relate and exchange it globally can only become effective if humanists engage with them and learn how to express their concepts, methods, and processes in detail, and in formalized ways .... Humanists on their own will not be able to harness the expressive power latent in the tools without an interdisciplinary collaboration with technologists and managers in which all parties have a common understanding of the possibilities of Semantic technologies and the structure and complexity of the humanists' discourse." (p. 255).

The project results have further implications for manuscript research and description in a digital and LOD environment. From the technical point of view, the MMM Portal provides a proof of concept for:

1. Provision of flexible access to the integrated data based on LOD publishing principles, such as the 5- and 7-star models, linked data browsing, and URI resolving;
2. Studying, analyzing, and reusing the data through a public SPARQL endpoint and APIs;
3. Intelligent user interfaces for semantic searching and exploring the global data (Marchionini, 2006) for more comprehensive views;
4. Applying seamlessly integrated data analysis tools over the global data integrated with the faceted search paradigm.

From the point of view of scholarship and metadata creation related to manuscripts, the project yields significant insights and recommendations for future work. An important lesson learned for the manuscript researchers on the team was the necessity of letting go of expectations of what data should do as opposed to what it actually could do. The initial analysis of the research questions and the discovery of semantic ambiguities in the structure of some questions, backed up by query failures such as querying the "popularity" of a manuscript, underscored how important it is for researchers, and by extension users, to understand the data model and the behavior of the data within the unified KG. For example,
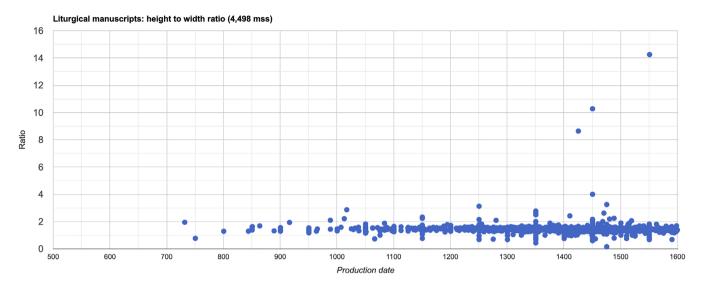
Liturgical manuscripts: height to width ratio (4,498 mss)

FIGURE 4 A visualization of height to width ratios of liturgical manuscripts in the KG

it may not be possible to align different vocabularies used in different datasets precisely, and errors in entity reconciliation may arise. The place concepts used in a dataset may be contemporary while historical places may be used in another. There may also be fundamental mismatches between the data models used.

A related concern for users is the danger of not understanding the "completeness" of the data, or in other words, how well the data represents all manuscripts. For the MMM data, the notion of "completeness" is further complicated by the data sources. MMM cannot claim universal coverage of manuscript data. The Oxford data only represents manuscripts in Oxford collections. The SDBM represents observations of manuscripts, many of which are duplicate observations of the same manuscript at different times of its existence. Bibale's focus on collections as opposed to individual manuscripts further skews the data picture. As a result, any analysis of the data has to be interpreted within the context of the collected data, not necessarily as a representation of a universal truth in the real world (Warren, 2018).

Rather than handicapping the project, the challenges presented by the aggregated data led the team to develop a system that could present the unique structures and relationships of data as transparently as possible. The result is a lighter user interface that illuminates rather than obscures the complicated pathways through the aggregated data, so that users eventually learn how to understand the semantic relationships and therefore how to ask better questions.

Other lessons learned that can be drawn from the MMM project for manuscript scholars and metadata specialists concern the implications for manuscript

description in a digital LOD environment. Traditional descriptive practices such as those commonly used in print catalogs favor unstructured prose. For manuscript metadata to be interoperable, however, it must be structured and dependent upon established authorities. For example, the provenance data in the Oxford records was presented descriptively in a note field, a common practice in traditional manuscript description. Although some elements had been consistently encoded, the full information embedded in the field could not be extracted for inclusion in MMM. The highly structured provenance data in the SDBM and Bibale might lack the narrative detail of the Oxford data, but it is more amenable to manipulation in a digital environment, much easier to extract, and more fully represented in MMM.

The experience of the data transformation process for provenance data thus suggests that in a SW environment descriptive manuscript data serves a different purpose than traditional manuscript description. The more that descriptions are standardized and embedded with external authorities in the description process, the more interoperability problems will be avoided during a transformation process (Hyvönen, 2010). In other words, where traditional manuscript description serves the human reader, manuscript description intended for use in a SW environment must also serve the needs of a computer's ability to read the data.

The research and lessons learned presented in this paper thus advance current methods and practices of manuscript description in a SW environment. The theoretical and practical insights into the computational representation of the history and provenance of manuscripts demonstrate the value of iterative and discursive

collaboration between those invested in manuscript research and those invested in semantic computing research. Diminishing the distance between these two poles enriches the intellectual enterprise of digital humanities research while at the same time building more robust and effective tools for research. While this project focused on manuscript studies, we firmly believe that our experience can be applied to future digital humanities work in any field or discipline.

ACKNOWLEDGMENTS

ORCID

Mikko Koho https://orcid.org/0000-0002-7373-9338
Toby Burrows https://orcid.org/0000-0002-0469-7584
Eero Hyvönen https://orcid.org/0000-0003-1695-5840
Esko Ikkala https://orcid.org/0000-0002-9571-7260
Kevin Page https://orcid.org/0000-0002-1668-6540
Lynn Ransom https://orcid.org/0000-0002-5231-3602
Jouni Tuominen https://orcid.org/0000-0003-4789-5676
Doug Emery https://orcid.org/0000-0002-5147-7736
Mitch Fraas https://orcid.org/0000-0003-3228-9876
David Lewis https://orcid.org/0000-0003-4151-0499
Andrew Morrison https://orcid.org/0000-0002-1074-8616
Guillaume Porte https://orcid.org/0000-0001-8656-8227
Emma Thomson https://orcid.org/0000-0002-5896-7922
Athanasios Velios https://orcid.org/0000-0002-4654-8675
Hanno Wijsman https://orcid.org/0000-0002-7236-2161

ENDNOTES

1 For example, most projects mentioned and detailed in Birnbaum et al. (2017).

2 https://sdbm.library.upenn.edu

3 https://tei-c.org/release/doc/tei-p5-doc/en/html/MS.html

4 https://www.w3.org/standards/semanticweb

5 http://cidoc-crm.org

6 https://www.ifla.org/publications/node/11240

7 http://dm2e.eu

8 https://github.com/bodleian/consolidated-tei-schema

9 https://github.com/bodleian/medieval-mss

10 https://medieval.bodleian.ox.ac.uk/

11 https://github.com/isl/Mapping-Memory-Manager

12 https://sdbm.library.upenn.edu/static/docs/SDBM_data_explanation2019.pdf

13 https://sdbm.library.upenn.edu/pages/SPARQL%20Data%20Model

14 https://sdbm.library.upenn.edu/sparql-space

15 http://erlangen-crm.org

16 https://sdbm.library.upenn.edu/pages/SDBM%20Place%20Authority

17 https://www.getty.edu/research/tools/vocabularies/tgn/about.html

18 http://vocab.getty.edu/doc/

19 https://github.com/mapping-manuscript-migrations/mmm-data-conversion

20 https://github.com/jiemakel/recon

21 https://www.go-fair.org/fair-principles/

22 https://zenodo.org/record/4440464

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

24 https://mappingmanuscriptmigrations.org

25 The public SPARQL endpoint: http://ldf.fi/mmm/sparql

26 https://www.w3.org/TR/sparql11-service-description/

27 https://www.w3.org/TR/void/

28 https://jena.apache.org/documentation/fuseki2/

29 https://nodejs.org/en/

30 https://expressjs.com

31 https://reactjs.org

32 https://redux.js.org

33 The Sampo-UI API is documented at https://mappingmanuscriptmigrations.org/api-docs

34 https://www.researchspace.com/

35 http://personal-research-domain-burrows.nodegoat.net/

36 The SPARQL query to retrieve the ratios of liturgical manuscripts in Yasgui: https://api.triplydb.com/s/yLQuKxaiM

37 https://diggingintodata.org

REFERENCES

Baierer, K., Dröge, E., Eckert, K., Goldfarb, D., Iwanowa, J., Morbidoni, C., & Ritze, D. (2017). DM2E: A linked data source of digitised manuscripts for the digital humanities. *Semantic Web*, *8*(5), 733–745. https://doi.org/10.3233/SW-160234

Barzaghi, S., Palmirani, M., & Peroni, S. (2020). Development of an ontology for modelling medieval manuscripts: The case of Progetto IRNERIO. *Umanistica Digitale*, 9, 117–140. https://doi.org/10.6092/issn.2532-8816/11187

Bellotto, A. (2020). Medieval manuscript descriptions and the semantic web: Analysing the impact of CIDOC CRM on Italian codicological-paleographical data. *Digital Humanities Quarterly*, 14(1) Retrieved from http://www.digitalhumanities.org/dhq/vol/14/1/000449/000449.html

Birnbaum, D. J., Bonde, S., & Kestemont, M. (Eds.). (2017). The digital middle ages: A *Speculum* supplement. *Speculum, 92* (S1). Retrieved from https://www.journals.uchicago.edu/toc/spc/2017/92/S1

Breitenfeld, A., Berger, F., Hong, M., Mackeprang, M., & Müller-Birn, C. (2018). Generating structured data by nontechnical experts in research settings. *i-com*, 17(1), 25–40. https://doi.org/10.1515/icom-2018-0005

Burrows, T. (2018). Connecting medieval and renaissance manuscript collections. *Open Library of Humanities*, 4(2), 32. http://doi.org/10.16995/olh.269

Burrows, T., Brix, A., Emery, D., Fraas, A. M., Hyvönen, E., Ikkala, E., Koho, M., Lewis, D., Myking, S., Ransom, L., Thomson, E. C., Tuominen, J., Wijsman, H., & Wilcox, P. (2020). *Linked Open Data vocabularies and identifiers for medieval studies*. Paper presented at Proceedings of Digital Humanities in Nordic Countries (DHN 2020). CEUR-ws.org. Retrieved from http://ceur-ws.org/Vol-2612/short5.pdf

Burrows, T., Emery, D., Fraas, A. M., Hyvönen, E., Ikkala, E., Koho, M., Lewis, D., Morrison, A., Page, K., Ransom, L., Thomson, E. C., Tuominen, J., Velios, A., & Wijsman, H. (2020). Mapping manuscript migrations knowledge graph: Data for tracing the history and provenance of medieval and renaissance manuscripts. *Journal of Open Humanities Data*, 6, 3. https://doi.org/10.5334/johd.14

Burrows, T., Holford, M., Lewis, D., Morrison, A., Page, K., & Velios, A. (2021). Transforming TEI manuscript descriptions into RDF graphs. *Graph data-models and Semantic Web technologies (GraphSDE) workshop 2019*. (Forthcoming).

Burrows, T., Pinto, N. B., Cazals, M., Gaudin, A., & Wijsman, H. (2020). Evaluating a semantic portal for the "Mapping Manuscript Migrations" project. *DigItalia*, 2, 178–185 Retrieved from http://digitalia.sbn.it/article/view/2643

Da Rold, O., & Maniaci, M. (2015). Medieval manuscript studies: A European perspective. *Essays and Studies*, 68, 1–25.

Doerr, M. (2003). The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3), 75–92. https://doi.org/10.1609/aimag.v24i3.1720

Doerr, M., Light, R., & Hiebel, G. (2020). *Implementing the CIDOC Conceptual Reference Model in RDF (V1.1)* (Report). Retrieved from http://www.cidoccrm.org/Resources/implementing-the-cidoc-conceptual-reference-model-in-rdf

Engels, R. (2020). Digital scholarship and medieval manuscripts: Access, technologies and potential. In B. A. Payer & A. Wall (Eds.), *Illuminating life: Manuscript pages of the middle ages*. The University of Guelph. https://hdl.handle.net/10214/21364

Frunzeanu, E., Robineau, R., & MacDonald, E. (2016). Biblissima's choices of tools and methodology for interoperability purposes. *CIAN-Revista de Historia de las Universidades*, 19, 115–132. https://doi.org/10.20318/cian.2016.3146

Gehrke, S., Frunzeanu, E., Charbonnier, P., & Muffat, M. (2015). *Biblissima's prototype on medieval manuscript illuminations and their context*. Paper presented at Proceedings of the First International Workshop Semantic Web for Scientific Heritage (SW4SH 2015). CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-1364/paper5.pdf

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space* (1st ed.). Morgan & Claypool Retrieved from http://linkeddatabook.com/editions/1.0/

Hyvönen, E. (2010). Preventing ontology interoperability problems instead of solving them. *Semantic Web*, 1(1–2), 33–37. https://doi.org/10.3233/SW-2010-0014

Hyvönen, E., Ahnert, R., Ahnert, S. E., Tuominen, J., Mäkelä, E., Lewis, M., & Filarski, G. (2019). Reconciling metadata. In H. Hotson & T. Wallnig (Eds.), *Reassembling the republic of letters in the digital age* (pp. 223–235). Göttingen University Press. https://doi.org/10.17875/gup2019-1146

Hyvönen, E., Ikkala, E., Tuominen, J., Koho, M., Burrows, T., Ransom, L., & Wijsman, H. (2019). *A Linked Open Data service and portal for pre-modern manuscript research*. Paper presented at Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019). CEUR-WS.org. Retrieved from http://www.ceur-ws.org/Vol-2364/

Hyvönen, E., Tuominen, J., Alonen, M., & Mäkelä, E. (2014). Linked data Finland: A 7-star model and platform for publishing and re-using linked datasets. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, & A. Tordai (Eds.), *The Semantic Web: ESWC 2014 satellite events* (pp. 226–230). Springer. https://doi.org/10.1007/978-3-319-11955-7_24

Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2021). Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. *Semantic Web*. Retrieved from http://www.semantic-web-journal.net/content/sampo-ui-full-stackjavascript-framework-developing-semantic-portal-user-interfaces-0

Klyne, G., & Lewis, D. (2020). *Exploring research questions through browsing: ResearchSpace for MMM* (Report). Retrieved from http://blog.mappingmanuscriptmigrations.org/wp-content/uploads/2020/12/Exploringresearch-questions-through-browsing_-ResearchSpace-for-MMM.pdf

Koho, M., Tuominen, J., Lewis, D., Ikkala, E., Heller, B., Thomson, E., Emery, D., Porte, G., Morrison, A., Velios, A., Wijsman, H., Hyvönen, E., Burrows, T., Ransom, L., Brix, A., Myking, S., Page, K., & Fraas, M. (2021). Mapping Manuscript Migrations knowledge graph (Version 2.2.0). *Zenodo*. https://doi.org/10.5281/zenodo.4440464

Le Bœuf, P. (2012). Modeling rare and unique documents: Using FRBRoo/CIDOC CRM. *Journal of Archival Organization*, 10(2), 96–106. https://doi.org/10.1080/15332748.2012.709164

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), 41–46. https://doi.org/10.1145/1121949.1121979

Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., & Van Harmelen, F. (2015). Semantic technologies for historical research: A survey. *Semantic Web*, 6(6), 539–564. https://doi.org/10.3233/SW-140158

Munby, A. N. L. (1960). *The dispersal of the Phillipps library*. Cambridge University Press.

Oldman, D., Doeer, M., & Gradmann, S. (2016). Zen and the art of linked data: New strategies for a semantic web of humanist knowledge. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new companion to digital humanities* (pp. 251–273). John Wiley and Sons. https://doi.org/10.1002/9781118680605.ch18

Page, K. R., Bechhofer, S., Fazekas, G., Weigl, D. M., & Wilmering, T. (2017). *Realising a layered digital library: Exploration and analysis of the live music archive through Linked Data*. Paper presented at 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 1–10). https://doi.org/10.1109/JCDL.2017.7991563

Ransom, L., Emery, D., Cawlfield, E., Heller, B., & Budisin, M. (2018). The new Schoenberg database of manuscripts: Creating an open-source tool for manuscript research and discovery. In M. J. Driscoll (Ed.), *Care and conservation of manuscripts 16: Proceedings of the Sixteenth International Seminar Held at the University of Copenhagen, 13th–15th April 2016*. Museum Tusculanum Press.

Riva, P., Doerr, M., & Žumer, M. (2009). FRBRoo: Enabling a common view of information from memory institutions. *International Cataloguing and Bibliographic Control*, *38*(2), 30–34 Retrieved from https://archive.ifla.org/IV/ifla74/papers/156-Riva_Doerr_Zumer-en.pdf

Sharpe, R. (2003). Titulus: Identifying medieval latin texts: An evidence-based approach. *Brepols*.

Tunkelang, D. (2009). Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *1*(1), 1–80.

Warren, C. N. (2018). Historiography's two voices: Data infrastructure and history at scale in the Oxford Dictionary of National Biography (ODNB). *Journal of Cultural Analytics*, *1*(2), 1–31. https://doi.org/10.22148/16.028

Wijsman, H. (2017). The Bibale database at the IRHT: A digital tool for researching manuscript provenance. *Manuscript Studies*, *1*(2), 328–341 Retrieved from https://repository.upenn.edu/mss_sims/vol1/iss2/10/

Zhitomirsky-Geffet, M., Prebor, G., & Miller, I. (2020). Ontology-based analysis of the large collection of historical Hebrew manuscripts. *Digital Scholarship in the Humanities*, *35*(3), 688–719. https://doi.org/10.1093/llc/fqz058

Zöllner-Weber, A., & Apollon, D. (2008). The challenge of modelling information and data in the humanities. In T. Hug (Ed.), *Media, knowledge & education—Exploring new spaces, relations and dynamics in digital media ecologies* (pp. 118–137). Innsbruck University Press. https://doi.org/10.26530/OAPEN_449459