Mapping Manuscript Migrations: Digging into Data for Researching the History and Provenance of Medieval and Renaissance Manuscripts: White Paper

Prepared by members of the "Mapping Manuscript Migrations" Project Team

Toby Burrows, Oxford e-Research Centre, University of Oxford
toby.burrows@oerc.ox.ac.uk (corresponding author)
Doug Emery, Schoenberg Institute for Manuscript Studies, University of Pennsylvania
Mitch Fraas, Schoenberg Institute for Manuscript Studies, University of Pennsylvania
Eero Hyvönen, Semantic Computing Research Group, Aalto University and University of Helsinki (HELDIG)
Esko Ikkala, Semantic Computing Research Group, Aalto University
Mikko Koho, Semantic Computing Research Group, Aalto University and University of Helsinki (HELDIG)
David Lewis, Oxford e-Research Centre, University of Oxford
Andrew Morrison, Bodleian Libraries, University of Oxford
Kevin Page, Oxford e-Research Centre, University of Oxford
Lynn Ransom, Schoenberg Institute for Manuscript Studies, University of Pennsylvania
Emma Thomson, Schoenberg Institute for Manuscript Studies, University of Pennsylvania
Jouni Tuominen, Semantic Computing Research Group, Aalto University and University of Helsinki (HELDIG)
Athanasios Velios, Ligatus, University of the Arts London
Hanno Wijsman, Institut de recherche et d'histoire des textes, Paris

Abstract

"Mapping Manuscript Migrations" (MMM) is a project funded between 2017 and 2020 by the Digging into Data Challenge. Its main goal is to combine data from several disparate sources about medieval and Renaissance manuscripts, and to use the aggregated data to explore a range of research questions about their history and provenance. The project took the data from three existing databases and turned them into Linked Open Data. This included mapping them to a newly developed unified data model, drawing on CIDOC CRM and $FRBR_{OO}$.

The aggregated data have been made available in several different ways. A copy of the dataset was published through the Zenodo repository. The data are hosted on the Linked Data Finland platform, from which they can be queried through a SPARQL endpoint or inspected directly. A semantic portal has also been implemented using the Sampo-UI user interface, through which

the 217,700 manuscripts and other entities can be searched, browsed, and analyzed, using a combination of filters and map-based visualizations.

A set of 25 research questions about manuscript history and provenance, provided by manuscript researchers, were used in designing the data model and the end-user perspectives for the semantic portal. They also formed the basis for an evaluation of the portal, in which the ability of the original three interfaces to the source datasets to answer the questions was compared with the new perspectives of the portal. This evaluation showed that the portal performed significantly better than the original interfaces and was capable of fully answering most of the questions.

Taken as a whole, the MMM project has demonstrated the value and potential of a Linked Open Data approach to representing, aggregating, and using data about medieval and Renaissance manuscripts in research, and has identified a number of important issues for the future of the approach. This paper examines the results of the project and the lessons learned from it.

Keywords

Linked Open Data; Medieval and Renaissance manuscripts; Data modelling and mapping; Provenance; Entity reconciliation

Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts: White Paper

1. Introduction

"Mapping Manuscript Migrations" (MMM) is a project funded between 2017 and 2020 by the Digging into Data Challenge of the Trans-Atlantic Platform. The main goal of the MMM project is to combine data from several disparate sources about medieval and Renaissance manuscripts, and to use the aggregated data to explore a range of research questions about manuscript history and provenance. The project took the data from three existing databases (the Schoenberg Database of Manuscripts, Medieval Manuscripts in Oxford Libraries, and Bibale) and turned them into Linked Open Data (LOD). This involved transforming them into RDF triples and mapping them to a newly developed unified data model, drawing on the CIDOC-CRM and $FRBR_{OO}$ ontologies. Vocabularies for the main classes of entity (manuscripts, actors, places, and works) were reconciled across the three data sources using a mixture of automatic and semi-automatic methods.

The aggregated data (nearly 22.5 million RDF triples) have been made available in several different ways. A copy of the dataset was published through the Zenodo repository. The data are hosted on the Linked Data Finland platform, from which they can be queried through a SPARQL endpoint or inspected directly. A semantic portal has also been implemented using the Sampo-UI framework (Ikkala et al., 2020), through which the 217,700 manuscripts and other entities can be searched and browsed, using a combination of filters and map-based visualizations. Results sets from the portal can also be downloaded as CSV files through a SPARQL query service like Yasgui.

Integral to the iterative design of the data model was a set of 25 research questions about manuscript history and provenance, provided by manuscript researchers. These questions were also used in formulating the filters for the semantic portal, and then formed the basis for an evaluation of the portal, in which the ability of the native interfaces to the source datasets to answer the questions was compared with that of the portal. This evaluation showed that the portal performed significantly better and was capable of fully answering most of the questions.

Taken as a whole, the MMM project has demonstrated the value and potential of a Linked Open Data approach to data about medieval and Renaissance manuscripts, and has identified a number of important issues for the future of such approaches. This paper examines the results of the project and the lessons learned from it.

2. Goals and scope

The MMM project addresses the proliferation of databases and other digital sources relating to medieval and Renaissance manuscripts. While there are numerous catalogues, lists, and digital collections available online, most are relatively limited in their coverage, and there are few methods of searching across these different sources. As a result, it is difficult and time-consuming for researchers to discover relevant information scattered in various places. Two notable exceptions have been the CERL Portal (discontinued from May 2020 because its technological basis had become obsolete) and Digital Scriptorium, which is currently working to re-develop itself for similar reasons. The MMM project was designed to explore next-generation solutions for combining data from multiple heterogeneous sources relating to medieval and Renaissance manuscripts, and especially to their history and provenance over the centuries of their existence. Of particular interest was the ability to transcend national boundaries, since many of the existing sources are purely local or national in scope. The project's approach to aggregating and presenting these heterogeneous datasets is based on the Linked Open Data framework.

The MMM project set itself the following specific goals:

- Combining and transforming data from at least three major public data sources;
- Exposing the combined data in a Linked Open Data environment;
- Implementing a software interface for browsing and searching the combined data;
- Visualizing the data to display relationships across time and space;
- Using the data to explore research questions relating to the history and provenance of manuscripts; and,
- Making the data and software available for reuse.

After nearly three years' work by a team of more than twenty people across four countries, the project is able to point to the following outputs. These are discussed in detail below.

- A unified Data Model for manuscript history and provenance, derived from the CIDOC-CRM and $FRBR_{OO}$ ontologies with some specific MMM additions, and based on the analysis of the data models used by the three source datasets and use cases gathered from scholars;
- A set of tools and pipelines for the transformation, mapping, and aggregation of the data from three source datasets, with a combined total of 262,000 records;
- The data from source datasets transformed to RDF, uploaded to the MMM triple store, and mapped to the MMM unified Data Model;
- Vocabularies for five main entity categories reconciled across the data sources: Manuscripts, Works, Events, Actors, Places;

- Linked Open Data vocabularies with unique identifiers for 217,700 manuscripts, 432,200 works, 4,700 places, 53,200 persons and organisations, and 915,300 events;
- A public SPARQL endpoint to the MMM triple store (containing almost 22.5 million RDF triples);
- A public portal to the MMM triple store offering browsing, searching, and visualizations of the aggregated data;
- Data export and downloads from the public portal, and a public data repository;
- A GitHub site containing tools, data, and documentation;
- More than 35 publications and presentations;
- A demonstration and advice for the CERL Portal Working Group, a significant industry body looking at best practice for libraries;
- Evaluation through a Focus Group of manuscript researchers and by the use of a set of 24 representative research questions;
- Enhancements to the three source datasets; and,
- Knowledge transfer between disciplinary groups, especially in the form of SPARQL expertise.

The MMM project has also identified a set of recommendations for improving the structure and reuse of manuscript provenance data in the future.

3. Use of Linked Open Data

The Linked Open Data (LOD) framework is integral to the design of the MMM project (Heath and Bizer 2011; Hyvönen 2012). A central hypothesis of the project is that this approach is able to express the kind of complex relationships which are found in humanities data, to enrich the data semantically by data aggregation and reasoning, and to allow for sophisticated and serendipitous discovery pathways and insights, more effectively than relational databases and marked-up documents can. In addition, the LOD framework is well-suited as an overarching structure for the interconnection of data sources which use different data models in the same knowledge domain. This kind of interconnection was a key goal of the MMM project.

The three datasets aggregated by the MMM project include two specialized relational databases which focus on the history and provenance of medieval and Renaissance manuscripts but use fundamentally different approaches to data modelling: Bibale (IRHT, 12,000 records) and the Schoenberg Database of Manuscripts (Schoenberg Institute for Manuscript Studies, 240,000 records). The former takes a single manuscript as the basic unit, and attaches the evidence of its history to that record. The latter focuses on an “observation” of a manuscript in a sales or auction catalogue or collection catalogue; there may be multiple observations of the

same manuscript, which are cross-linked to show this relationship.[1] The third dataset – Medieval Manuscripts in Oxford Libraries – consists of 10,000 XML documents which each describe a single manuscript using the Text Encoding Initiative (TEI) markup for a manuscript description.

The MMM data are expressed in RDF triples, governed by a data model constructed from the CIDOC-CRM and $FRBR_{OO}$ ontologies with some additional MMM-specific extensions. The CIDOC-CRM components relate mainly to the description of the manuscript as a physical object and to its event-centred history (sales, gifts, and other ownership events). The $FRBR_{OO}$ components relate mainly to the intellectual content of the manuscript: the works and authors of the texts carried by the physical object. The MMM extensions express specific characteristics relevant to manuscripts and manuscript-related events. The data model is discussed in more detail below.

Each of the data sources maintains and deploys an extensive set of vocabularies for the main MMM entity classes, especially Persons, Organizations, and Places. Following the standards and best practice in the Linked Open Data world (https://www.w3.org/standards/semanticweb/), their authority records for these entities often contain references to the main general vocabularies used for these types of names, notably the Virtual International Authority File (VIAF) for persons and organizations, and GeoNames and the Getty Thesaurus of Geographic Names (TGN) for places. For the titles of Works, both Bibale and the Oxford catalogue maintain an authority list with some external references; the Schoenberg Database, on the other hand, tracks generic titles rather than works in the specific sense employed by FRBR. The existence of a significant number of vocabulary-based references in the source datasets was another incentive for using the Linked Open Data framework in the MMM project. These references were critical to the process of automatically reconciling and linking data from different sources relating to the same specific entity. Other entities – especially Works and Manuscripts themselves – were reconciled by semi-automatic means; this mainly consisted of identifying possible matches programmatically and then asking manuscript researchers and librarians to examine them one-by-one to find the definite matches.

4. Data modelling

A major element in the MMM project was the time and effort devoted to the development of the unified data model which is at the heart of the affordances offered by the MMM aggregated dataset and the various interfaces to it. A Modelling Group, consisting of the

[1] For a more detailed explanation of the SDBM data model, see https://sdbm.library.upenn.edu/static/docs/SDBM_data_explanation2019.pdf.

project's Semantic Web and LOD experts and a representative of the manuscript researchers and curators, met weekly for more than twelve months to inspect and analyse the incoming data and to identify the entity classes and properties which were required to express the structures of the three different data sources. The scoping of this data model was done in a pragmatic way; the aim was to reflect the scholarship embedded in the sources and identify their common features, rather than carrying out a theoretical review of the entire domain of medieval and Renaissance manuscript research based on the published literature.

The source datasets each reflect an earlier process of in-depth analysis and model construction. In two cases (Bibale and the Schoenberg Database of Manuscripts) the provenance and history of manuscripts were the focus of this process. In the third case (the Bodleian catalogue) the focus was somewhat broader: the detailed description of manuscript characteristics, contents, and histories. For MMM, the entity classes and properties identified from the three sources were compared with CIDOC-CRM and $FRBR_{oo}$, to see how many of them could be expressed in terms of these ontologies. About one-third of them could not, and for these a specific MMM entity class or property was created. Table 1 shows the sources of the different elements in the MMM data model. The properties column includes multiple occurrences of the same property under different entity classes; it excludes four generic properties from the OWL, Dublin Core Terms, and SKOS ontologies, as well as four geographical properties from the Getty Vocabulary Program and WGS84 Geo Positioning ontologies. The full schema is included as Appendix 1.

Source	**Entity Classes**	**Property**
CIDOC-CRM Erlangen	18	80
$FRBR_{oo}$	4	3
MMM-specific	11	65
Total	33	148

Table 1: MMM Data Model – sources

The main classes in this model are those for manuscripts, manuscript collections, the texts carried by a manuscript, persons, organizations, places, and events. Several of these were relatively straightforward mappings to CIDOC-CRM: Place (E53), Person (E21), Group (E74), and Actor (E39). For Work (F1) and Expression (F2), the $FRBR_{oo}$ definitions and elements were used. A mixture of CIDOC-CRM and $FRBR_{oo}$ classes was required to cover the range of different types

of events: F27_Work_Conception, F28_Expression_Conception, E12_Production, E10_Transfer_of_Custody, and the more generic E7_Activity. Two additional types of events specific to the history of manuscripts were also defined: mmms:ActorActivity and mmms: ManuscriptActivity.

The entity class which required the most detailed modelling was F4 Manifestation Singleton (from $FRBR_{OO}$), used for single manuscripts in the sense defined by the FRBR ontology: "physical objects that each carry an instance of F2 Expression, and that were produced as unique objects, with no siblings intended in the course of their production." In FRBR terms, a manuscript is a unique manifestation of an expression of a work. In both Bibale and the Oxford catalogue, the basic record describes one manuscript. The Schoenberg Database, on the other hand, consists of records which are observations of a manuscript at a point in time. Where two or more observations have been linked, the linking "manuscript record" is mapped to the F4 Manifestation Singleton class; where an observation has not been linked, it is also mapped to the F4 class.

The F4 class in MMM had no less than 29 different properties, 18 of which were specific to the MMM data model. They included mmms:last_known_location and mmms:phillipps_number, used respectively to consolidate data about the last known location of a manuscript (since current location was only known in a minority of cases) and to enable matching of manuscript records by their use of the same number from the collection of Sir Thomas Phillipps. Among the MMM-specific entity classes were a series of specific physical attributes for manuscripts: mmms:Folios | mmms:Columns | mmms:Lines | mmms:DecoratedInitials | mmms:HistoriatedInitials | mmms:Miniatures. The data described with these properties could be modelled with CIDOC-CRM and $FRBR_{OO}$ properties and entities or retrieved using relevant queries, but often such processes result in long chains in the graph or complex queries. CIDOC-CRM properties such as 'P56 bears feature' could be used instead of those proposed here, but establishing more specialised properties such as 'DecoratedInitials' serves central aspects of the discourse of this project while maintaining compatibility with the CIDOC-CRM.

One area of considerable discussion in the Modelling Group related to the current location and ownership of manuscripts, an area of obvious interest to users of the MMM project's outputs. This kind of information is implicit but clear in a library-based catalogue like that of the Bodleian. In Bibale, the current library shelf-mark is usually recorded in the title field of a manuscript record and includes the place where that library is located. This information is not specifically present in the Schoenberg Database, although it can often be inferred from the most recent observation about a specific manuscript. A similar inference also needs to made in Bibale, but the reasoning involved is more straightforward; the The uncertainty and risk of error inherent in such inferential processes would need to be made clear to the user, and expressed

in some complex form in the data model. The Modelling Group, with the advice of the manuscript researchers and curators, decided to model instead the concept of a "last-known location". This was expressed in three separate properties of a manifestation singleton, reflecting the last-known location in as many of the three data sources as contained a specific manuscript. An overall property (mmms:last_known_location) could then be calculated, using a fixed order of preference: Bodleian, Bibale, Schoenberg Database. This instantiated, as far as possible, the level of certainty or uncertainty about the manuscript's current location and ownership.

A set of 25 research questions assembled by the project's manuscript researchers with an Oxford focus group were used intensively in the development and testing of the data model. These are discussed in more detail below. In the second stage of the project, once the first version of the user interface had been implemented, some additional properties were included as a rest of the process of testing. They included mmms:manuscript_author and mmms:manuscript_work, two shortcut properties designed to create a direct relationship between manifestation singletons and the works and authors they contained – alongside the more complex indirect relationship prescribed by $FRBR_{OO}$.

Working through the research questions led to refinements of the incoming data. It was clear, for example, that entries in the Schoenberg Database for sales and gifts did not have place information associated with them. This would have made it impossible to answer several of the research questions, even though the data model specified the relationships between these kinds of events and the locations in which they occurred. As a result, the Schoenberg Database entries for sales catalogues from specific organizations (such as Sotheby's) were enriched with place information. Once the enhanced data were uploaded and mapped to the MMM data model, answers to these kinds of questions could be populated.

The data model is a major intellectual output of the MMM project. Other attempts to model manuscript descriptions using CIDOC-CRM and $FRBR_{OO}$ have been more focused or more limited in scope. The Biflow ontology developed for a catalogue of medieval Tuscan bilingual texts covers the linguistic and literary aspects of the manuscripts as well as their "material and historical characteristics" (Mancinelli et al. 2019). The ontology deployed for the Biblissima observatory's prototype "Medieval Manuscript Illuminations and their Context" focuses on the illustrations described in two databases: Mandragore and Initiale (Gehrke et al. 2015). In comparison, the MMM data model covers all aspects of manuscript research, including provenance, history, physical description, and textual content.

5. Data transformation

The data transformation pipeline for the three source datasets follows four basic steps:

- Transform the source data into RDF triples, using the native structure of the dataset itself, and expose them for harvesting;
- Harvest and upload the transformed data to a central store, hosted by the Semantic Computing Research Group at Aalto University, and combine them as Turtle input files;
- Map the uploaded RDF triples to the MMM unified Data Model, using automated SPARQL CONSTRUCT statements;
- Validate the resulting output; and,
- Reconcile, as far as possible, the vocabularies for the main entity classes in order to link instances of the same entity occurring in two or more data sources.

In the case of Bibale and the Schoenberg Database of Manuscripts, the initial transformation is from a relational database to RDF triples. In the case of the Bodleian Library, the initial transformation is from TEI-encoded XML documents, which involves a more complex form of pre-processing (Burrows et al. 2020). The first step in the Bodleian Library's workflow is to identify those parts of the TEI schema which are needed to answer the research questions of the MMM project. An xQuery script is used to extract these parts and copy them into a simplified XML document. It also creates URIs for each included entity. This simplified XML output is then mapped to classes and properties of the CIDOC-CRM and $FRBR_{OO}$ ontologies using the 3M mapping tool (Oldman, Theodoridou and Samaritakis 2010). The Bodleian Library's XML authority files are handled as separate datasets following the same method. Manuscript instances are then integrated with the authority records via corresponding URIs. The records include references to URIs from external authorities such as VIAF, GeoNames, TGN, Gemeinsame Normdatei (GND), and WikiData.

The work done to transform the Bodleian Library's documents for the MMM project has demonstrated that it is possible to extract TEI-encoded manuscript data in a form which can be expressed as RDF, loaded to a graph database, incorporated into a Linked Data environment, and retrieved using SPARQL queries. But the nature of some of the TEI markup – and especially the lack of encoding for various components of the narrative <provenance> statements – means that the RDF representation cannot include all the relevant semantic content from the catalogue records. In the Bodleian Library's catalogue itself, the keyword search function can still find occurrences of (for example) a bookseller's name, even though these names have not been encoded. Replicating this functionality in the RDF environment would mean either re-encoding the TEI files in a more thorough and structured way or developing additional scripts to parse, extract, and transform provenance information which is currently presented in unencoded narrative statements within a <provenance> element. These options were not possible within the time frame of the MMM project, but would be suitable tasks for a follow-on project.

The MMM data integration pipeline is not reliant on live synchonization across different datasets. Each of the three source datasets is continually being updated with new content, however, so the pipeline is designed to be repeatable. At the moment, each new upload must be initiated manually rather than run automatically. This at least means that any future changes to the sources will be noticed immediately when the transformation process is run, instead of an automatic failure, though it does also mean that updates must be scheduled and resources allocated.

6. Exploring the data

Two main user interfaces are available for exploring the aggregated data. A SPARQL endpoint maintained by the Semantic Computing Research Group at Aalto University makes it possible to run SPARQL queries against the entire dataset. This approach requires a level of technical familiarity which is probably not found amongst most manuscript researchers' skills. There are benefits to working with the 'raw' data when conducting detailed investigations, perhaps after initial enquiries through visualisations (below); digital humanities practitioners may well have learnt how to construct SPARQL queries using a tutorial like that published by the Programming Historian (Lincoln 2015). The MMM project implemented a weekly SPARQL training session for project staff who did not already have this kind of expertise, through knowledge transfer from colleagues at Aalto University. These sessions demonstrated the value of SPARQL for constructing complex questions to take advantage of the full extent of the MMM unified data model. Queries using the Yasgui interface can be recorded as URLs, which can then be used to re-run the original query. Some examples can be seen in the MMM SPARQL Tutorial.

In most cases, however, manuscript researchers are likely to use, at least initially, the public MMM Semantic Portal, which provides an interface to the data through the Sampo-UI framework developed by the Semantic Computing Research Group at Aalto University: https://seco.cs.aalto.fi/tools/sampo-ui Sampo-UI is used across four other semantic portals in Finland and Norway, as well as a further five Finnish portals currently in development. It consists of a client built on various JavaScript libraries, especially React and Redux, and a backend API which converts a request into a SPARQL query using a set of query templates and configurations, runs the query against a preconfigured SPARQL endpoint, processes the SPARQL results with a preconfigured result mapper, and returns them in JSON or CSV format: https://github.com/SemanticComputing/sampo-ui

An important goal of the MMM portal was to enable users to browse the entire combined dataset, rather than relying solely on a keyword search interface. Accordingly, the portal provides five "perspectives" on the data, i.e., avenues for browsing using the main classes of

entities: Manuscripts, Works, Events, Actors, and Places. Each of these perspectives displays an extensive range of data about each entity of this type. The Manuscripts perspective, which is the most fully-developed and is expected to be the starting-point for most users, displays a table listing 24 different types of information about each of the more than 217,700 manuscripts in the MMM dataset, including: production place, production date, last known location, authors, works, languages, owners, collections, transfer of custody dates and places, and so on.

Each manuscript also has its own "landing-page" accessible from the table, which lists all the information about that manuscript under 25 headings. There is a link from each manuscript to its record in the Linked Data Finland platform, with its complete set of classes and properties, as well as an option to connect to the Yasgui SPARQL query service and download this information in the form of a CSV spreadsheet.

Most users will be aiming to identify groups or sets of manuscripts which meet a particular set of criteria, such as a combination of their contents, their language, their place of production, their last known location, their physical characteristics, or their former or current owners. An extensive set of filters enables several of these criteria to be combined in complex ways, e.g., "illuminated manuscripts of works by St Augustine in Latin produced in fourteenth-century France, with the United States as their last-known location." The result set can also be downloaded as a CSV spreadsheet through the Yasgui SPARQL query service.

The results of such queries – and also the entire dataset – can be visualized against three maps. The first shows the places where manuscripts were produced, as far as these are known and documented. The second shows the last-known locations of the manuscripts, while the third shows their migrations from production to last-known location in the form of arcs between these two places. Because data for the sequence of ownership are often unclear and lacking in usable dates and places, it is not currently possible to visualize each step of a manuscript's travels over the centuries, though this information can be seen and downloaded in a table. So the visualizations only cover, at most, three data points: a manuscript, its place of production, and its last known location. Even at this level of reduction, the full "migrations" visualization appears heavily overloaded in its initial form. It is easy, however, to zoom in until individual arcs start to appear and can be clicked on to see the details. Nevertheless, the full visualization has its own value and impact, since it conveys the basic message of the MMM project: that many thousands of medieval and Renaissance manuscripts have travelled widely across Europe and the world in the centuries since their initial production.

Similar but less extensive approaches are available for the other four perspectives. The tabular presentations have fewer columns, while the visualizations are more limited: a map showing the location of provenance and transfer events, a map showing places associated with persons

and organizations, and a map showing the places mentioned in the dataset. For these other classes of entities too there are "landing-pages" with links to the Linked Data Finland record and to the Yasgui download option.

The initial version of the MMM portal was presented at a workshop during the Digital Humanities conference in Utrecht in July 2019, attended by 18 people from the manuscript research, digital humanities, and library communities. They were given a schedule of tasks and asked to give detailed feedback as they worked through this process. The results were then used to improve both the functionality of the portal and the help information and documentation.

7. Reusability of data and software

An important goal for the MMM project was to ensure the reusability of both the aggregated data and the software deployed by the project. A GitHub site was set up to make the software available and has been used by three of the project partners to share scripts and software: https://github.com/mapping-manuscript-migrations Its main components are:

- The Sampo-UI software developed by Aalto University and used for the MMM Semantic Portal (written in JavaScript);
- A Docker container for populating a Fuseki triplestore with the MMM Knowledge Graph;
- Scripts for uploading the RDF files produced from the source datasets and transforming them to the MMM Data Model;
- The Bodleian Library's xQuery scripts for producing RDF from its TEI-XML documents; and,
- The Ruby script developed by the Schoenberg Institute for Manuscript Studies for modifying the Oxford TEI-XML files with additional provenance and acquisition information.

The GitHub site also contains extensive project documentation, including the Data Model and a SPARQL tutorial, as well as some of the data, notably the initial RDF output from the Bodleian Library's transformation process.

The aggregated MMM data have been published in the Zenodo repository. Version 1.1.0 (14 February 2020) of the data – amounting to about 1.25 GB in total – is available for download: https://doi.org/10.5281/zenodo.3667486 The data are made available as RDF Turtle files. There is one file for each of the three source datasets, containing the transformed and mapped source data in the form of RDF triples, and including the reconciled instances of Manuscripts, Works, and Actors. Also deposited are a separate "Places" file, which contains the RDF triples for the reconciled places, and a "Schema" file containing the data model.

The data are available in other ways too. MMM provides a public SPARQL endpoint from which the dataset can be queried: http://ldf.fi/mmm/sparql The linked data are served by the Linked Data Finland platform hosted by Aalto University: http://www.ldf.fi/dataset/mmm/ Result sets from searches in the MMM Semantic Portal can be exported in the form of CSV spreadsheets through the Yasgui public SPARQL query interface: https://yasgui.triply.cc/#

The MMM data are made available for reuse under a CC BY-NC 4.0 license: https://creativecommons.org/licenses/by-nc/4.0/ Two main reuse cases are envisaged, both of which would be applicable to researchers studying such subjects as the history of medieval and Renaissance manuscripts, the history of collecting and collections, and the transmission and dissemination of classical, medieval, and Renaissance texts. The first case would cover the whole dataset; there were sixteen downloads from the Zenodo repository in the first two months of availability. The Oxford e-Research Centre has loaded a copy of the entire dataset into a different software environment – ResearchSpace (developed by MetaPhacts and the British Museum) – and is currently configuring a new interface, which will include a network visualization of the data (Oldman and Tanase 2018). The second case applies to a selection of the data, identified through the portal or a SPARQL query. One of the authors (Burrows) is downloading a sub-set of the data relating to a specific manuscript collector (Sir Thomas Phillipps) for import into a nodegoat database of Phillipps manuscripts, using CSV spreadsheets as the transport mechanism (Burrows 2017).

These exports could also be used to extend or add to the existing visualizations. While timelines are used for filtering in several of the perspectives, for example, and charts can be constructed from the owner data in the Manuscripts perspective, exporting a selection of the data into a specialized visualization software environment would be feasible. The same would apply for constructing network diagrams or life-path visualizations (Sankey diagrams) for manuscripts, neither of which is available in the MMM portal itself.

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

Renaissance studies. The MMM identifiers can be referenced from other LOD-compatible vocabularies and portals, and MMM entities can also be pulled into SPARQL queries in other services using the SERVICE keyword. It would also be possible, for example, to build an annotation layer on to the MMM dataset using a SPARQL-based API.

8. Knowledge transfer and outreach

The expertise available to the MMM project covered several very different areas: manuscript studies, digital humanities, and Semantic Web research. An important component of the project was to ensure an effective level of knowledge transfer between these areas, beginning with an initial orientation to manuscript provenance research for the Semantic Web experts on the project. The set of 25 research questions developed in the initial stage of the project proved vital to this work, since they could be used to explain the specific elements which needed to be covered in the MMM data model as well as the kinds of functionality which manuscript researchers would expect from the MMM portal.

Knowledge transfer within the project focused on the growth of RDF, SPARQL, and Linked Open Data expertise among manuscript researchers and librarians as well as some technical staff who had not had previous exposure to this kind of knowledge. They included staff supporting Bibale and the Schoenberg Database of Manuscripts, who were able to transform their databases into RDF triples and make them available for harvesting. At the University of Oxford, this kind of knowledge was transferred internally between the Oxford e-Research Centre and the Bodleian Libraries' Digital Library Service. A particularly successful initiative was a weekly SPARQL session, which began in July 2019 and was still running ten months later. It enabled librarians and technical staff from the project to learn and practice the use of SPARQL queries against the MMM dataset, with the advice and assistance of the Semantic Web experts from Aalto University. The set of research questions provided a good basis for designing these SPARQL queries.

Knowledge transfer outside the project began with a focus group held in the Bodleian Library in November 2017, attended by twelve manuscript researchers ranging from doctoral students to senior academics. In response to an introduction to the project's goals and methodologies from project staff, the participants gave some very useful ideas about the kinds of questions and functionality that a manuscript provenance service might be expected to provide. A workshop was then held at the Digital Humanities Conference at Utrecht in July 2019 for a group of eighteen manuscript researchers and digital humanities experts, which provided valuable initial feedback on the functionality of the first version of the MMM portal and on the work done by the project to that stage more generally.

In February 2020, the MMM project was presented to an expert group of librarians, curators, and digital humanities specialists from national and regional libraries across Europe. They were advising the Council of European Research Libraries (CERL) on the future of its increasingly obsolete Manuscripts Portal, which will be decommissioned in mid-2020. Useful exchanges of information took place with representatives of the Bibliothèque nationale de France about the future of the Biblissima platform, and with representatives of German libraries developing a new national portal for medieval and Renaissance manuscripts: https://handschriftenportal.de/projekt/ Two similar presentations were made in November 2017 and March 2019 to meetings of manuscript cataloguing experts from the University of Oxford, the University of Cambridge, and the British Library.

Outreach for the MMM project was largely carried out through a dedicated Twitter account with more than 330 followers, and a Web site with a blog, as well as through more than forty accepted conference presentations. A further nine were scheduled for events which were postponed or cancelled due to the coronavirus epidemic. These presentations, like the publications emanating from the project, were targeted at conferences and journals representing each of the different disciplinary areas involved in the project: manuscript studies, medieval and Renaissance studies, digital humanities, and Semantic Web research.

9. Improvements to source datasets

The MMM project has been careful to define and maintain a clear and persistent relationship between the RDF triples created by the project and the original source data. In the MMM portal, users can always refer back to the original datasets via links provided in each entity's "landing page". The MMM data can also be filtered by source, for direct access to a source's dataset if required. In this sense, the RDF data created by the MMM project are, effectively, a supplementary layer to the source information.

The transparent relationship between the source datasets and the MMM data underscores the role of MMM as an aggregator of data rather than as a data management system, though an unanticipated but welcome outcome of the project has been its ability to help managers of the original datasets to identify problems. Because data correction is not part of the MMM transformation process, weaknesses, inconsistencies, and errors in the datasets become clear in search results, alerting dataset managers that something needs to be fixed at their end. In this way, MMM enables managers to clean and enrich their data. For instance, in the Schoenberg Database and Bibale, hundreds of personal and institutional names have been corrected for authority control, resulting in a rich and as yet untapped record of names associated with manuscript production and trade. But value has also been added to these datasets. In the Schoenberg Database, locations were added to records for sellers – particularly

those firms and other organizations which held sales and auctions. When harvested by MMM, this information provides locations for sales events, which can enhance the MMM visualizations. The Schoenberg Database has also been able to mount a public SPARQL endpoint, which enables a search of the pre-transformation RDF version of the data.

Additionally, more than 2,000 of the Oxford TEI files have been updated with structured provenance information relating to previous collection owners that can now be pulled into the RDF transformation. This was done by forking a copy of the relevant XML files from the Bodleian Library's GitHub site, and running a Ruby script which added a standard statement with appropriate TEI markup to record the acquisition of each manuscript in a collection from its previous owner. The new versions of the files were then re-loaded to the GitHub site and checked by Bodleian manuscript librarians before replacing the previous version. From the GitHub site, the new version could then be harvested into the MMM transformation pipeline, as well as being pushed to the Bodleian's own Web catalogue.

10. Organizational issues

The nature of the funding process, with each of the four partner institutions funded directly by their national funding agency, meant that each partner started the project at different times: two in July 2017, one in December 2017, and one in April 2018. This made necessary some re-thinking of the initial project plan and timetable. The schedule for transforming the datasets, in particular, had to be altered in order to start with the available partner's data, rather than arranging the order of the datasets by more technical criteria. The end dates for the partners have been less of a problem, though here too one partner will not finish their funding until October 2020 while the others are all finishing in mid-2020.

The spread of resources across the four partners also required careful coordination, since the number of project staff, the nature of their skills and expertise, and the roles they were expected to play in the project all varied considerably. The Aalto University team had the primary responsibility for transforming the data, hosting the project's triple store, and implementing the portal's user interface. Two partners had specialist Linked Open Data and Semantic Web expertise – Aalto University and the Oxford e-Research Centre at the University of Oxford – but they started work about six months apart. It was especially necessary to ensure that sufficient expertise of this kind was available across the course of the project. Three of the partners (the Schoenberg Institute for Manuscript Studies, the Institut de recherche et d'histoire des textes, and the Bodleian Libraries) were responsible for providing and transforming the data, and for applying knowledge of manuscript curation and research to the design of the data model and the user interface.

The project worked mainly through online meetings using the BlueJeans software hosted by the University of Pennsylvania. Two specialized groups were formed, focusing on data modelling and the user interface respectively. The Modelling Group contained all the project's Linked Open Data experts, together with a librarian, and met weekly for eighteen months from February 2018 to June 2019. It developed and refined the unified data model in an iterative process, and formed the main channel to the work being done at Aalto University to set up and host the transformation pipelines and the aggregated Linked Open Data. The Users Group met fortnightly from August 2018 to November 2019. Its membership consisted of the manuscript researchers and librarians on the project, together with the user interface design expert from Aalto University. It focused mainly on the development and testing of the Sampo-UI interface. Meetings of the full project team were also held monthly, beginning in May 2018.

Two face-to-face meetings of the project team were held, beginning with a two-day kick-off meeting at the University of Oxford in September 2017, which focused on planning and scoping the project. A second three-day meeting was held in Helsinki in April 2019. This reviewed the work done to date, discussed and resolved the detailed issues arising, and defined and agreed on a work plan for the final stages of the project. The success of the Helsinki meeting was crucial to completing the project successfully. In retrospect, a similar face-to-face meeting in mid-2018 might have been a valuable opportunity for reviewing the data model and planning the implementation of the user interface.

11. Evaluation through research questions

The project team developed a set of 25 research questions to guide its progress. Some of these questions were elicited from the initial focus group of manuscript researchers, while others were contributed by members of the MMM project team or taken from a list produced by the French project Biblissima ("Requêtes intéressantes"): https://doc.biblissima.fr/ontologie-biblissima - méthodologie Some of these questions were specific, e.g.: "Which manuscripts containing texts by Ramon Llull were sold in the 19th century?" Others were more generic, e.g.: "How many illuminated manuscripts were in a particular collection?" The initial use of these questions was in developing the MMM unified data model; they were later used to test and refine the filtering and searching capabilities of the user interface.

The research questions were also used to evaluate the MMM Portal and its Linked Open Data framework. For this process, each question was tested first against the three source datasets individually. While each of these sources provides a relatively sophisticated interface, in almost every case it proved difficult to answer the questions fully (Table 2). At best, the user was presented with a partial answer to the question, often in the form of a broader list of results

which had to be scanned manually to identify relevant items. Some questions could not be answered at all using the source databases alone (8 in Bibale, 8 in Oxford, 6 in Schoenberg).

In the MMM Portal, on the other hand, a majority of the questions (17 out of 25) could be answered readily 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 explored further by running queries against the MMM SPARQL endpoint. This approach was able to provide full answers to such questions as: "Which collectors bought manuscripts from Wilfrid Voynich? Where were the collectors located? What do we know about the kind of manuscripts he sold, and their earlier histories?" The full list of research questions, with the results of their testing, is given in Appendix 2.

	Bibale	Oxford	Schoenberg	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

One of the specific but complex research questions used as an example in the original funding application 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 catalogue. In Bibale, it is impossible to run a query on transactions of a specific period, while in the Oxford catalogue the list of people can be filtered by role (e.g., owner) but not by place, whether this is place of birth, place of death, or place of residence – depending on the definition of "French." The Schoenberg Database does make it possible to identify people linked to France with life dates after 1598, and then view the individual entries linked to them. But this will not cover people linked to specific places within France, since the 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 1598 involves adding one of the timeline filters to find persons born after (say) 1550. 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.

12. Future directions and lessons learned

The MMM project has demonstrated the value and effectiveness of a Linked Open Data approach to aggregating provenance data for medieval and Renaissance manuscripts. The sophisticated data model and the transformed data have been made available for reuse, while the MMM portal shows how searching, filtering, and visualizing can be successfully applied to answer complex research questions. An obvious desideratum for the future is to transform and incorporate data from a wider range of sources, with the aim of both increasing the number of manuscripts covered and adding to the information available for the 217,700 manuscripts currently represented.

The project has identified several areas where future work would be valuable. The first is the development of specialist Linked Open Data vocabularies for medieval studies. The project was able to make effective use of such general vocabularies as VIAF for Actors and TGN for Places, since these were used in the source datasets. The history of manuscripts focuses on the nineteenth and twentieth centuries as much as on the medieval period, which reduces the need to identify medieval names. But there are still a significant number of people (as authors and manuscript owners), organizations (especially religious houses), works, and places which do not appear in the more general vocabularies. There are various existing lists and databases of medieval names which could be transformed into Linked Open Data to enrich discovery services and knowledge graphs in this field. A typical example is the database "Monasteries in the Netherlands until 1800: a census", which contains records for about 750 religious houses: https://www2.fgw.vu.nl/oz/monasteries/index.php It is searchable but not downloadable, and does not include LOD identifiers, either for the houses themselves or for references to other datasets.

The MMM project also identified gaps in the provision of authoritative vocabularies for the people and institutions associated with book production, the book trade, and book collecting. While the CERL Thesaurus contains "forms of imprint places, imprint names, personal names and corporate names" connected with the book trade, its scope is limited to "material printed before the middle of the nineteenth century." VIAF is limited to names associated with book publishing as supplied by the world's national libraries; the Schoenberg Institute for Manuscript Studies is ineligible to participate in this programme, despite offering to contribute.

Discovery and linkage of manuscripts specifically would benefit greatly from the existence and use of unique LOD identifiers. Manuscripts are generally identified by their owner, the collection, and a unique shelf-mark or catalogue number (e.g., British Library, Egerton MS 8546), but frequent inconsistencies in formatting shelf-marks and collection names, even within the institution itself, can make it hard to match data relating to the same manuscript. Frequent changes in ownership and in owners' names, even for institutions, can also cause problems in reconciling shelf-marks and catalogue numbers for the same manuscript. The MMM project used Phillipps numbers as one way of linking data about the same manuscript, and this matched nearly 9,000 manuscripts. The ISMI (International Standard Manuscript Identifier) initiative has been established with the aim of defining a unique manuscript identifier, but there has been only limited progress so far (Cassin 2018).

Another area of future work lies in improving the way in which provenance histories for manuscripts are recorded. The MMM project worked with two specialized and sophisticated provenance-oriented databases (Bibale and the Schoenberg Database of Manuscripts). Though they provided MMM with a substantial amount of well-structured data, their data models were quite different from each other. Library catalogues based on the MARC record, on the other hand, usually give provenance information in an unstructured note. The MMM project did not attempt to incorporate this kind of data, which would have required the use of text analysis and entity recognition techniques. There is some scope to incorporate more structure into this type of annotation.

The Bodleian Library's TEI-XML documents use the <provenance> tag to encode information about manuscript ownership. This tag is geared towards the kind of narrative histories and notes about provenance evidence found in traditional printed manuscript catalogues, and its contents are largely unstructured, with the exception of the names of persons mentioned in the narratives. As a result, the MMM project found it difficult to extract anything more than a generic event from this kind of data. The project did not have time to design and implement a programme of text analysis and entity recognition techniques for this unstructured information, though this approach would be worth applying in the future.

A possible way forward for improving the structure of provenance data, especially in TEI-encoded catalogues, might be found in the work of the Linked Art project, which is developing a general schema for art history, using CIDOC-CRM: https://linked.art/model/index.html The MMM project is currently working to produce guidelines for a more structured approach to the TEI encoding of manuscript history and provenance statements. The aim is to find a framework for recording and encoding such data which is sufficiently well-structured to map to data models like that of MMM. This will increase

the specificity and richness of the kind of computer processing, analysis, and visualizations implemented by MMM.

The MMM project has learned a great deal about using Linked Open Data in humanities research. The process of developing the unified Data Model was complex and iterative, involving extensive dialogue between technical experts and manuscript researchers. For more than twelve months, weekly meetings of the MMM Modelling Group worked to design a model which reflected the knowledge of scholars in this field. This model was tested initially against the research questions assembled by the manuscript researchers, and the results were fed into further rounds of model development. The model did not try to reflect the entire body of knowledge found in the published literature of manuscript research. Instead, its scope was determined pragmatically, with only the classes of entities and relationships identified in the source datasets being included.

The project also produced insights into different methods of querying and navigating the RDF triples, and the resulting benefits. As well as using and refining the portal framework, this also involved direct access to the triple store through the SPARQL endpoint. One of the results was a better understanding of the trade-offs between pre-built user interfaces and the direct use of SPARQL, and an appreciation of the real benefits in using the query language to discover new insights into the data. Articulating a research question in SPARQL can appear complex because one has to be precise in articulating the query, but that complexity in query articulation is really a reflection of complexity in the data. It needs a manuscript scholar's familiarity and judgement to decide how "correct" the combination of query parameters is or should be. The data and the query cannot be objective, so it is important for scholars to be able to interrogate and judge the subjectivity in the data and the query. This, in the context of the MMM project, is what SPARQL "forces" upon us.

More generally, the MMM project offers some insights into the relationship between digital products based on Linked Open Data and scholars' expectations. Instead of an opaque interface in the form of a simple Google-style search box, which conceals the complexity of the raw data beneath it, the MMM portal and its SPARQL endpoint are designed to open up the data for close inspection and exploration. The portal enables a user to browse most of the data points for each type of entity directly, and provides a "landing page" for each entity to show all the available data relating to it. SPARQL queries enable a user to interrogate the RDF triples directly, using the full features of the data model. The Linked Data Finland service and the Zenodo copy of the dataset make the underlying RDF triples available for export and reuse.

This approach helps to reveal the complexity and provisional quality of the data and the data structures. It also reveals the added value of both the data modelling and the data

transformation pipeline. The dataset itself adds value and novelty and is the result of an intellectual process – analysing the source data and the field of knowledge itself. The design of the MMM outputs is intended to strike a balance between reflecting existing knowledge and encouraging new types of research questions. The data model is sufficiently broad in scope and detailed in its coverage to limit the extent to which exploration is prescribed. A user can simply browse through the portal and combine filters at random, and then see if the result is a meaningful pattern or not. This is the equivalent of deriving research questions from the data, rather than always approaching the data with pre-compiled research questions. The aim is to deploy the Linked Open Data framework in ways that reflect the iterative nature of humanities research and preserve the richness of its evidence base.

Acknowledgements

The Mapping Manuscript Migrations project was funded under Round 4 of the Trans-Atlantic Platform's Digging into Data Challenge (2017-2020). The four project partners -- University of Oxford (Oxford e-Research Centre and Bodleian Libraries), the University of Pennsylvania (Schoenberg Institute for Manuscript Studies), the Institut de recherche et d'histoire des textes (IRHT-CNRS), and Aalto University (Semantic Computing Research Group) -- were funded respectively by the following agencies: Economic and Social Research Council (UK), Institute of Museum and Library Services (US), Agence nationale de la recherche (France), and the Academy of Finland.

The project also benefited from work done as part of the OXLOD Project, funded by the University of Oxford's IT Capital Plan as part of the GLAM Digital Strategy.

The authors wish to acknowledge the contributions of the following former and current project staff: Pip Willcox (University of Oxford), Benny Heller (University of Pennsylvania), Nicole Bergk-Pinto, Antoine Brix, Mahaut Cazals, Alexandre Gaudin, Synnøve Myking, Pierre-Louis Pinault, and Guillaume Porte (Institut de recherche et d'histoire des textes)

Competing Interests

The authors have no competing interests.

Author contributions

Toby Burrows T.B.
Doug Emery D.E.
Mitch Fraas M.F.
Eero Hyvönen E.H.
Esko Ikkala E.I.
Mikko Koho M.K.
David Lewis D.L.
Andrew Morrison A.M.
Kevin Page K.P.
Lynn Ransom L.R.
Emma Thomson E.T.
Jouni Tuominen J.T.
Athanasios Velios A.V.
Hanno Wijsman H.W.

Conceptualization T.B., E.H., L.R., and H.W.
Data curation D.E., M.F., M.K., D.L., A.M., E.T., J.T., A.V., and H.W.
Formal analysis D.E., M.K., D.L., K.P., E.T., and J.T.
Funding acquisition T.B., E.H., L.R., and H.W.
Methodology D.E., M.K., D.L., K.P., E.T., and J.T.
Project administration T.B., E.H., K.P., L.R., and H.W.
Software D.E., E.I., M.K., D.L., A.M., J.T., and A.V.
Supervision T.B., E.H., K.P., L.R., J.T., and H.W.
Visualisation E.I.
Writing – original draft T.B.
Writing – review & editing E.H., E.I., M.K., K.P., L.R., J.T., A.V., and H.W.

Works cited

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

Burrows, Toby, Athanasios Velios, Matthew Holford, David Lewis, Andrew Morrison, and Kevin Page. 2020. "Transforming TEI Manuscript Descriptions into RDF Graphs," GraphSDE proceedings (forthcoming)

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

Gehrke, Stefanie, Eduard Frunzeanu, Pauline Charbonnier, and Marie Muffat. 2015. "Biblissima's Prototype on Medieval Manuscript Illuminations and their Context." In: SW4SH 2015: Semantic Web for Scientific Heritage 2015: Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC Conference, Portorož, Slovenia, June 1st, 2015, Arnaud Zucker, Isabelle Draelants, Catherine Faron Zucker, and Alexandre Monnin (eds). http://ceur-ws.org/Vol-1364/paper5.pdf

Heath, Tom, and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space* (Synthesis Lectures on the Semantic Web: Theory and Technology). Palo Alto: Morgan & Claypool. http://linkeddatabook.com/editions/1.0/

Hyvönen, Eero. 2012. *Publishing and using cultural heritage linked data on the Semantic Web*. Palo Alto: Morgan & Claypool.

Ikkala, Esko, Eero Hyvönen, Heikki Rantala and Mikko Koho. 2020. "Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces." May, 2020. Submitted. https://seco.cs.aalto.fi/publications/2020/ikkala-et-al-sampo-ui-2020.pdf

Lincoln, Matthew. 2015. "Using SPARQL to access Linked Open Data," *The Programming Historian*. https://programminghistorian.org/en/lessons/retired/graph-databases-and-SPARQL

Mancinelli, Tiziana, Antonio Montefusco, Sara Bischetti, Maria Conte, Agnese Macchiarelli, and Marcello Bolognari. 2019. "Modelling a Catalogue: Bilingual texts in Tuscan Middle Ages (1260-1430)." Poster, Digital Humanities 2019, Utrecht University, 8-12 July 2019. https://dev.clariah.nl/files/dh2019/boa/1219.html

Oldman, Dominic, and Diana Tanase. 2018. "Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace," in: The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II, ed. Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa et al. (Berlin: Springer, 2018), pp. 325-340.

Oldman, Dominic, Maria Theodoridou, and Georgios Samaritakis. 2010. "Using Mapping Memory Manager (3M) with CIDOC CRM." Version 4g. http://83.212.168.219/DariahCrete/sites/default/files/mapping_manual_version_4g.pdf

Appendix 1: MMM Schema

Class/Property	Range
Namespaces:	
PREFIX dct: http://purl.org/dc/terms/	
PREFIX ecrm: http://erlangen-crm.org/current/	
PREFIX frbroo: http://erlangen-crm.org/efrbroo/	
PREFIX mmms: http://ldf.fi/mmm/schema/	
frbroo:F1_Work	
dct:source	mmms:Database
mmms:data_provider_url	URL
skos:altLabel	string
skos:prefLabel	string
frbroo:F27_Work_Conception	
ecrm:P4_has_time_span	ecrm:E52_Time-Span
ecrm:P7_took_place_at	ecrm:E53_Place
dct:source	mmms:Database
skos:prefLabel	string
frbroo:R16_initiated	frbroo:F1_Work
mmms:carried_out_by_as_author	ecrm:E21_Person
mmms:carried_out_by_as_possible_author	ecrm:E21_Person
mmms:carried_out_by_as_commissioner	ecrm:E21_Person
mmms:carried_out_by_as_editor	ecrm:E21_Person

frbroo:F2_Expression, ecrm:E33_Linguistic_Object.	
ecrm:P72 has language	string
dct:source	mmms:Database
mmms:data_provider_url	URL
skos:altLabel	string
skos:prefLabel	string
frbroo:F28_Expression_Creation	
ecrm:E12_Production	
ecrm:P4_has_time_span	ecrm:E52_Time-Span
ecrm:P7_took_place_at	ecrm:E53_Place
ecrm:P108_has_produced	frbroo:F4_Manifestation_Singleton
dct:source	mmms:Database
skos:prefLabel	string
mmms:carried_out_by_as_commissioner	ecrm:E21_Person
mmms:carried_out_by_as_illuminator	ecrm:E21_Person
mmms:carried_out_by_as_printer	ecrm:E21_Person
mmms:carried_out_by_as_scribe	ecrm:E21_Person
frbroo:F4_Manifestation_Singleton	
ecrm:P128_carries	F2 Expression
ecrm:P3_has_note	string
ecrm:P43_has_dimension	"mmms:Width, ..."
ecrm:P45_consists_of	mmms:Material
ecrm:P46_is_composed_of	frbroo:F4_Manifestation_Singleton
ecrm:P46i_forms_part_of	ecrm:E78_Collection

ecrm:P51_has_former_or_current_owner	ecrm:E21_Person
ecrm:P52_has_current_owner	ecrm:E21_Person
ecrm:P70i_is_documented_in	mmms:Source
dct:source	mmms:Database
mmms:catalog_or_lot_number	string
mmms:data_provider_url	URL
mmms:entry	URL
mmms:manuscript_author	ecrm:E21_Person
mmms:manuscript_record	URL
mmms:manuscript_work	frbroo:F1_Work
mmms:phillipps_number	string
mmms:shelfmark_arsenal	string
mmms:shelfmark_barocci	string
mmms:shelfmark_bnf_hebreu	string
mmms:shelfmark_bnf_latin	string
mmms:shelfmark_bnf_nal	string
mmms:shelfmark_buchanan	string
mmms:shelfmark_christ_church	string
owl:sameAs	URI
skos:altLabel	string
skos:prefLabel	string
ecrm:E52_Time-Span	
ecrm:P81a_end_of_the_begin	datetime
ecrm:P81b_begin_of_the_end	datetime
ecrm:P82a_begin_of_the_begin	datetime
ecrm:P82b_end_of_the_end	datetime

skos:altLabel	string
skos:prefLabel	string
E10 Transfer of Custody / E7 Activity	
skos:prefLabel	string
ecrm:P11_had_participant	ecrm:E21_Person
ecrm:P28_custody_surrendered_by	ecrm:E21_Person
ecrm:P29_custody_received_by	ecrm:E21_Person
ecrm:P30_transferred_custody_of	ecrm:E21_Person
ecrm:P3_has_note	string
ecrm:P4_has_time-span	ecrm:E52_Time-Span
ecrm:P70i_is_documented_in	Source
ecrm:P7_took_place_at	ecrm:E53_Place
mmms:data_provider_url	URL
mmms:observed_manuscript	frbroo:F4_Manifestation_Singleton
mmms:ownership_attributed_to	ecrm:E21_Person
dct:source	mmms:Database
skos:prefLabel	string
ecrm:E78_Collection	
ecrm:P51_has_former_or_current_owner	
ecrm:P92i_was_brought_into_existence_by	
ecrm:P93i_was_taken_out_of_existence_by	
mmms:collection_location	
mmms:collection_type	
mmms:data_provider_url	

mmms:external_url	
mmms:institution_literal	
mmms:location_literal	
mmms:source_agent	
mmms:source_date	
mmms:source_type	
dct:source	
skos:altLabel	
skos:prefLabel	
mmms:Source	
mmms:source_date	
mmms:source_agent	
mmms:source_type	
mmms:external_url	
mmms:location_literal	
mmms:institution_literal	
mmms:data_provider_url	
mmms:external_url	
dct:source	mmms:Database
mmms:Source_Type	
skos:prefLabel	string
ecrm:E53_Place	
gvp:placeTypePreferred	currently string, actually an AAT concept

gvp:broaderPreferred	ecrm:E53_Place
wgs84:lat	decimal
wgs84:long	decimal
dct:source	mmms:Database / Getty TGN / Geonames
owl:sameAs	URI
mmms:data_provider_url	URL
ecrm:P89_falls_within	ecrm:E53_Place
ecrm:E21_Person / ecrm:E74_Group / ecrm:E39_Actor	
ecrm:P98i_was_born	ecrm:E67_Birth / ecrm:E66_Formation
ecrm:P100i_died_in	ecrm:E69_Death / ecrm:E68_Dissolution
ercm:P3_has_note	string
mmms:gender	
mmms:religion	
mmms:biblissima_id	
mmms:data_provider_url	URL
dct:source	mmms:Database
skos:prefLabel	
ecrm:E67_Birth / ecrm:E69_Death	
ecrm:P4_has_time-span	ecrm:E52_Time-Span
ecrm:P7_took_place_at	ecrm:E53_Place
skos:prefLabel	
ecrm:E63_Beginning_of_Existence / ecrm:E64_End_of_Existence	
ecrm:P4_has_time-span	ecrm:E52_Time-Span
ecrm:P7_took_place_at	ecrm:E53_Place
dct:source	mmms:Database

skos:prefLabel	
ecrm:E66_Formation / ecrm:E68_Dissolution	
ecrm:P4_has_time-span	ecrm:E52_Time-Span
ecrm:P7_took_place_at	ecrm:E53_Place
skos:prefLabel	
ecrm:E57_Material	
skos:prefLabel	string
mmms:Height / mmms:Width	
ecrm:P90_has_value	
ecrm:P91_has_unit	mmms:Millimetre
mmms:Folios / mmms:Columns / mmms:Lines / mmms:DecoratedInitials / mmms:HistoriatedInitials / mmms:Miniatures	
ecrm:P90_has_value	integer
mmms:Database	
mmms:data_provider_url	

Appendix 2: Research Questions used by the MMM Project

MMM Research Question	Bibale results – public interface	Oxford results – public interface	SDBM results – public interface	MMM results - portal	MMM portal - comments
[A1] How many manuscripts from pre-1600 produced in European countries survive?	Currently impossible. One cannot search by date span, and one cannot search by place either (not by country, let alone by continent).	Can't filter by "Europe" (only by a specific country). Each century has to be filtered separately – can't aggregate "pre-1601". Can show "how many MSS have a production event in England in the 14thcentury?"	In advanced search, limit Manuscript Date to terminating at 1601. Then limit these results with Place facet. This search returns 90,185 Entries. Then open "Manuscript" facet to see all MS records associated with place.	In "Manuscripts" perspective: (1) Filter by "Production Place = Europe" (produces 71,718); (2) Set "Production Date" time slider to 1601. Result is 67,231.	In step (1), the figure shown next to "Europe" in the left-hand panel is 82,971 - this is the count of all Production Places, which is greater than the count of all Manuscripts

[A2] How many manuscripts survive that contain Spanish texts written in gothic rotunda were produced in Castile for an abbey or convent? Then show me those which were owned during the nineteenth century by English private collectors; Then show me those which are now owned by an institution in North America.	Currently impossible. This is a complex question anyway.	Not possible. You can filter the list of manuscripts by language (Spanish) and then by place of origin (Spain - not region) - which results in 14 manuscripts. But you can then only examine the results individually to analyse their provenance and history.	Can browse language facet for Spanish, which gives 2027 results. These results can be narrowed down by items in the Place facet linked to "Castile" (9) and "Castilla et Leon (13). From there individual entries must be analyzed for data related to 19th century owners.	In the "Manuscripts" perspective: (A) filter by language = Spanish. (B) Then filter by production place = Castile. (C) Then, as a proxy for filtering by script, filter by date of production, e.g. 1100-1450. This produces 34 results. (D) Then look through the list of owners to identify medieval Spanish abbeys/convents (3), 19th-century private English owners (5), and 20th-century North American institutions (2). (E) Repeat this process for manuscripts produced in places in Castilla-Leon or Castilla-La Mancha (since these are separate from Castile in the TGN hierarchy).	(A) It is not possible to filter manuscripts by script. (B) It is not possible to filter manuscripts by type of owner (private / public / institutional / religious house etc.). (C) An alternative approach would be to start from the "Agents" view, and filter for "type of owner = group" + active in Castile. You can then see the manuscripts and works associated with these institutions as owners. You would then need to look at each of these manuscripts individually to see their subsequent

					owners.
[A3] What French collectors purchased manuscripts since the end of the Wars of Religion (after 1598)? Where are their manuscripts now?	Currently impossible. One can currently not run a query on transactions of a specific period (eg. after a specific year).	You can filter the list of "People" by role (e.g., "Owner, signer, or donor". But you cannot further filter this list by country - whether this is place of birth, place of death, or place of residence.	This question can't easily be answered. You can search for SDBM Names that are linked to France and have life dates after 1598, but then you'd have to run separate queries to find people linked to places nested within France (there's currently no way to search for France and all of its children within the same query). Once you found all of these names, you could then view the entries linked to them and determine last known locations, but this would be very time consuming.	In the "Actors" perspective: (A) Limit the type of actor to E21: Person. Then (B) filter by Activity Location = France. This gives 1,698 results. Then (C) filter the "Birth" timeline to births after 1550. This gives 572 results. Then (D) look at the individual manuscripts associated with each person to see their subsequent history.	You need to sort the resulting list by "Role" to distinguish manuscript/collection owners from authors of works. At step (C) you could also filter the "Death" timeline for deaths after 1600. This produces 475 results. If you combine the births and deaths timeline limits, the result is 475 people. There is no obvious way of filtering the list of manuscripts associated with these people.

[B1] How many manuscripts containing texts by Ramon Llul were sold in the 19th century?	One can run a query on "Lullus" and filter the word only in the field "Contenu indicatif (auteur, titre)". Then one can look (by simply opening the files) if one of these was sold in the 19th century. For the moment this is not so complicated, since there is only one manuscript with the word "Lull" in the field "Contenu indicatif (auteur, titre)". Then one can also run the query "lullus" and open the three Works in the result list to see if any of them appears in a manuscript. This, however, is not the case for the moment.	You can browse or search to find all manuscripts with Ramon Llull as an author. But their provenance/history can only be analysed by looking at each manuscript individually.	Not easily answered. You can search for entries that list Llull as an author with provenance dates in the 19th century, but few entries list specific provenance dates (7 entries result from the above search). A better way would be to search for source dates in the 19th century, but in the current SDBM interface it isn't possible to combine an author facet with a range of source dates within the same search.	In the "Manuscripts" perspective: (A) search for Author "Llul" - this finds "Ramon Llull, with 343 results. (B) Filter for Transfer of Custody date = 1800-1900, with four results.	The "Transfer of Custody" results are understated, given that many provenance events are only categorized as "Activity". But you can't tackle this question from the "Events" view, since that only lists manuscripts, not authors or works.

[B3] Who collects manuscripts with texts by Ramon Llul?	One can run such a query in several steps (see B1), but for the moment there are no results in the data.	You can browse or search to find all manuscripts with Ramon Llull as an author. But their provenance/history can only be analysed by looking at each manuscript individually.	You can quickly find all entries with texts by Ramon Llull using a basic search. From those results, you can browse the Provenance facet to see a list of all associated provenance agents. However, this information isn't easily exported.	In the "Manuscripts" perspective: (A) search for Author "Llul" - this finds "Ramon Llull, with 343 results. (B) Look through the list of Owners to see which owners have been associated with these manuscripts.	There isn't a way to filter on current or past owners, or to sort the owners by date. But you could sort the results lists on either the "Event" or "Transfer of Custody Date" column to get some indication of recent or current ownership.
[B4] How many times do texts by Ramon Llul's appear with texts by Albertus Magnus in the same manuscript?	One can run such a query in several steps (see B1), but for the moment there are no results in the data.	It is possible to do a keyword search of the Manuscript records for a combination of the two names, e.g. "Albertus AND Lull" or "Albertus OR Lull". Using "Llull" in this search produces no results, however, even though that is the preferred form in the Author record - because the search is on the text of the Manuscript entry and that uses	There are 16 entries that contain texts by both Llull and Albertus Magnus in the SDBM.	In the "Manuscripts" perspective: (A) search for Author = "Llul" - this finds "Ramon Llull", with 343 results. (B) Look through the list of manuscripts to see which of them also contain Albertus as an author.	If you select two or more authors in the "Author" filter, the results are combined as an OR operation (not as AND). It is not possible to search the Llull result set to find those containing "Albertus". Sorting the Llull result set by Author is of limited help, since a manuscript may contain authors other

		"Lull", not on the Authors file which uses "Llull".			than Llull or Albertus, and only the first author is displayed in the summary list of results.
[C1] What was the most popular text by a medieval author in France in the 17th Century?	Currently impossible. This is a complex question anyway. One cannot search by date span or by place.	Not possible. You can't even sort the list of "works" by the number of times they occur in Oxford manuscripts.	Difficult to answer since source date is not searchable in a date range	The closest approximation is probably as follows. (1) In the "Manuscripts" perspective, limit the "Transfer of Custody date" to 17th century. This produces 373 results. (2) Then limit the "Transfer of Custody place" to France. This produces 30 results. (3) Then sort by Author or Work. Five medieval authors are represented, each with one Work.	(A) This does not cover the full range of ownership Events. (B) The "Events" view can be filtered by date range (17th century) and by place (France), but the result is a list of manuscripts, not of works or authors. (C) The "Works" view cannot be filtered by Place or Event (only by language or production date).

[C2] Did Sir Thomas Phillipps own a 13th-century Bible with historiated initials?	A query for "Phillipps" and filter on Collections, shows us the file of Phillipps library. At the moment there are some 500 manuscripts described. In the list, however, we only see shelfmarks, so we will have to open them one by one to see if there is a 13th-century bible among them. Alternatively, we can do a general query on the word "bible" and filter only the field "Contenu indicatif (auteur, titre)" and only on the object "Livre (Exemplaire)". This gives a list of about 140 manuscripts. In this list of shelfmarks we will have to open them one	(1) You can easily identify all manuscripts formerly owned by Phillipps by selecting his name from the list of Persons. But the resulting list of manuscripts cannot be filtered or analysed in any way - except by inspecting each entry individually. (2) An Advanced Search for "Phillipps" can be filtered for century of origin and presence or absence of decoration. But it cannot be refined by title/contents/works.	Yes: there are 3 entries that describe 13th century bibles that contain historiated initials with Phillipps as a provenance agent.	In the "Manuscripts" perspective: (1) Filter for "Phillipps" in Owner - result is 8,752 manuscripts. Then (2) search in "Work" filter for "bible" - result is 130 manuscripts. Then (3) search in "Historiated Initials" filter for "minimum = 1". The combined result is 10 manuscripts. Sorting on the "Historiated Initials" column shows a range from 1 to 150 initials.	If you search in the "Work" filter for "biblia" instead, this produces 1 additional manuscript with historiated initials.

	by one to see if there is a 13th-century bible among them.				
[F1] Combien de manuscrits enluminés se trouvent dans une collection particulière? (volumétrie)	There is no simple way to search this. But any collection present in Bibale can be opened and will include a list of associated books. The files of these books should then be opened one by one to see if they are illuminated or not.	An Advanced Search for "Phillipps" can be filtered for presence or absence of decoration in the resulting list of manuscripts.	You can easily limit results to entries owned by a particular person/institution, and then further limit by any number of physical characteristics.	In the "Manuscripts" perspective: select a specific owner in the"Owner" filter. Then combine this with a selection of "minimum = 1" in one or all of the filters for "Miniatures", "Decorated Initials", or "Historiated initials". For Thomas Phillipps as owner, there are 555 manuscripts with at least one decorated initial, 198 with at least one historiated initial, and none with miniatures. There are 14 with both types of initials.	You can also filter on "Collection" instead of "Owner". For the Phillipps Catalogus, this produces results of 413 with decorated initials, 141 with historiated initials, and 10 with both types of initials.

[F3] Qui sont les donateurs et les propriétaires d'une collection?	Any collection present in Bibale can be opened and will include a list of associated books, as well as associations to donators, though one should also look in the file of the person to look for associations with donors. The collection will also show links to one or several owners.	The Oxford catalogue can be filtered to see a list of former owners and/or donors across the whole of the modern Oxford collection. But there is no way of filtering for former owner against specific sub-collections (e.g., colleges, named Bodleian collections).	SDBM doesn't include donor relationships (though a general chain of ownership can be established within entries, showing a direct transfer of custody from one provenance agent to the next). Some entries may state explicitly that someone donated the manuscript to another collection, but this information isn't searchable. In most cases, the best you can do is show a generic association between two provenance agents.	In the "Manuscripts" perspective, select one (or more) of the Collections in the "Collections" filter. Then look at the list of owners displayed in the "Owners" filter. This shows the number of MSS in that collection which were also associated with specific owners. This information can also be visualized using the "Chart" functionality.	

[F4] Faire des recherches par sujet, par technique, par langue, par artiste voire par pigments (plus d'encre d'or, argent et pourpre) dans une collection.	Direct searches on these fields cannot be operated for the moment, but any collection present in Bibale can be opened and will include a list of associated books. These books can then be opened one by one to see if there are data on language, technique, etc.	(1) The Manuscripts list can be filtered by type of material, by language, and by the presence or absence of decoration. (2) The People list can be filtered by role = scribe, or role = artist, with associated lists of manuscripts against each person.	SDBM doesn't have pigment, technique or subject searching, but you can search by language and artist easily.	In the "Manuscripts" perspective, select one (or more) of the Collections in the "Collections" filter. You can then inspect the other filters to see various characteristics of the MSS in that collection: material, size, presence of initials and decoration, and language. You can browse the list of Works contained in these MSS, but not by subject.	
[F5] Particularités d'une collection (sujet, technique, lieu de production etc.) ? Quelles en sont les lacunes ? Quelles en sont les dominantes ?	Browse to find a manuscript; in the record look at the list of associations with persons and with collections. These two lists can both be ordered by the dates and by the places that are mentioned in the respective columns on the right, but these are the dates and places of	Browse "Manuscripts", then filter by "Collection". Select one Collection, and then filter by "Origin" (to see a list of countries and regions, with numbers of MSS) or by "Century" (to see a list of centuries with numbers of MSS).	Limit search results to those Entries with X Name as Provenance Agent. From this search results page, browse the Place facet to see a list of all places of production associated with those Entries (Numerical Sort=number of entry appearances). Browse	In the "Manuscripts" perspective, select an owner or a collection. Then use the "Production Places" map visualization to see the distribution of their MSS by place of production. For the distribution of production dates, you can sort the list of MSS by "Production date".	

	the association, not of the origin of the manuscript; one cannot filter by those.		Manuscript Date facet similarly.		
[F6] Vie d'une collection, vie d'un livre enluminé ?	Any collection present in Bibale can be opened and will include a list of associations (to books, to persons, to works, etc.). This will give the user information about the live of the collection. When one has an account with administrator rights, one can visualise this in a diagram (a feature that should become available to all in the future).	There are no visualizations, though it might (remotely) be possible to export the raw TEI files from Github and process them into some visualization software.	There's no way for a user to create visualizations within the database, but you can export any search results as a .csv, and then use different software to create a visualization with that file. You could search for all of the entries linked to a certain person/institution, or all of the entries describing the same manuscript, to create the visualization.	In the "Manuscripts" perspective, select one of the collections listed in the "Collections" filter. You can then visualize the life of this collection through the "Production Places" map, the "Last Known Locations" map, and the "Migrations" map. You can get the same visualizations for a specific individual MS as well by searching on its "Label" details in the "Label" filter. The "landing-page" for a specific individual MS also provides a full history of the life of that MS in tabular, rather than graphical, form.	

[F7] Quels manuscrits sont probablement perdus ?	Any list of manuscripts in Bibale can be displayed in shelfmark order. Thus by hand one can see which names are indeed a shelfmark and which are another designation (sale so and so, inventory such and such, etc.). But those that do not have a current shelfmark are of course not necessarily lost. One can	The Oxford catalogue only covers MSS known to be in an Oxford library today.	This is impossible to answer in the SDBM interface.	In the "Manuscripts" perspective, you can sort by "Last Known Location" to identify those MSS which have no "Last known location."	
[F8] Quel manuscrit a été vendu et n'est pas identifié au sein d'une collection à l'heure actuelle ? (catalogue de vente)	See F7. In the result list all names starting with "Vente" (sale) have no current shelfmark and are only known by their last sale.	The Oxford catalogue only covers MSS known to be in an Oxford library today.	This is difficult to answer in the SDBM interface. The best we can do is search for Entries that represent recent sales, and then isolate individually those that have no known buyer or more recent observation.	The best you can do is to browse through those MSS with no "Last known location" (see F7), and identify those with at least one "Provenance Event".	

[G1] Quelles copies d'un texte sont enluminées ?	Not easy to do at the moment in Bibale. Technically, a query for a text could show you the list of manuscripts after which you can open these files one by one to see if there are data on the illumination, but there are few text files and few data on illumination.	The list of Manuscripts can be filtered by the presence or absence of decoration; each manuscript record would then have to be inspected individually. The list of Works cannot be filtered in this way.	SDBM entries record the count of miniatures and/or initials, but not simply whether a manuscript is illuminated or not. SDBM also doesn't have a work concept. The best you can do is search for a specific title, and then limit to those entries whose miniatures or initial fields aren't blank.	In the "Manuscripts" perspective, search for a specific Work using the "Work" filter. The resulting list of MSS can then be filtered or browsed for the presence of miniatures, decorated initials, or historiated initials.	This cannot be done through the "Works" perspective, however.
[G2] Quelle position occupe une copie dans l'histoire de la transmission d'un texte ? Y a-t-il des exemplaires uniques des oeuvres ?	This is not possible in Bibale (or can only be done by selecting each work separately to see how many MSS are linked to it, but even that is tricky because simply calling up a list of all the works is currently not possible)	Can only be done by selecting each of the 10,987 Works separately to see how many MSS are linked to each specific Work.	When browsing the Title facet, sort Titles by Numerical Sort, then navigate to page 310 of the results, where Titles appearing in only one Entry begin to appear in the results. This will return results related to Entries, not MS Records.	You can browse the list of "Works" to see which works only have a single MS attached. This does mean going through more than 400,000 entries for Works, however. You can also sort the list of Works by manuscript, but this still involves inspecting all the entries for Works.	The most effective way of doing this is through a SPARQL query.

[G4] Quelles sont les versions existantes d'une oeuvre ? Qui a fait une traduction française d'un texte ancien ? Quand ?	A query of a name of an author or a title of a work with a subsequent filter on the object "Work" can easily lead you to any work, after which the file of the work will show if there are associations to other works signifying the one being a translation or a reworked version of the other.	The list of Works can be browsed to see versions of the same Work in different languages - but only inasmuch as they are contained in Oxford manuscripts. The names of the translators and the dates of the translations are not normally specified.	SDBM doesn't have works. You can search by various title names to gather different versions of the same text, including versions in different languages. Depending on the entry data, this would allow you to determine the date of the first appearance of a text in a certain language.	In the "Works" perspective, search for a specific work using the "Title" filter (together with the "Author" filter if necessary). The resulting list can be filtered by language to see different translations of that Work, and the MSS in which they appear. The translator's name, if recorded, will appear in the "Possible author" column of the results list.	
[5] Quelles sont les différentes publications existantes [manuscript copies] d'un texte ? (date, lieu de production, personne(s) responsable(s) etc.)	Browse to find the specific work you are looking for; then in the record of this work one will find in the list of "associations" the list of all the manuscripts it is linked to.	Browse by "Works", select a specific Work, and see a list of all Oxford MSS containing that Work.	Browse via the Title facet, or search on the Title field. Then limit results to entries produced in a certain location (using the Place facet or field). This will return results related to Entries, not MS Records.	Browse the "Works" perspective (or search for a specific work). Select a Work to view its "landing page". This lists all MSS of that Work, together with their dates and places of production.	

[H1] How many manuscripts were produced in Northern Italy and/or Lombardy?	One can do a general browse on the word “Lombardie” (for example) and then filter by the field “lieu” and by the object “livre”.	Can filter places of production by some regions (Flanders, Dalmatia, etc.) but not by regions within most countries.	Search on the Place field, or browse via Place facets. Searching on only “Northern Italy” or “Lombardy” returns 6,467 Entries. If you search on all of the regions of Northern Italy, 6,538 (due to SDBM nesting errors). One can open “Manuscript” facet to see all ms records associated with place.	In "Manuscripts" perspective: (1) Filter on "Production Place" by traversing the hierarchy to reach "Lombardy" - then tick box. Result is 702. (2) Clear the result from Step 1, then filter on "Production Place" by traversing the hierarchy to reach "Northern Italy. Result is 944. (The alternative is to use the "Bounding Box" option under Filter Options, and draw a rectangle around the geographical areas. This gives a result in the order of 6,186 manuscripts.)	(A) You cannot combine two different places in the "Production Place" filter. (B) In Step (1), the figure shown against "Lombardy" in the left-hand panel is 728 rather than 702. The figure for "Northern Italy" is 944. (C) The Bounding Box can only be rectangular, and cannot follow the contours of geographical regions. (D) "Northern Italy" is not a hierarchical region in the TGN vocabulary, so Step (2) will not pick up regions like the Veneto which are in Northern Italy but not in Lombardy.

[H2] How many manuscripts were produced in the Low Countries?	One can obtain results by general browses on the modern countries “Netherlands”, “Belgium” and “Luxemburg” and then filter by the field “lieu” and by the object “livre”. Then one could add similar browse on the region “Nord-Pas-de-Calais” to obtain more accurately the historic region of the “Low Countries”.	Can filter places of production by “Flanders”	Search on the Place field, or browse via Place facets. Searching on Low Countries returns 9,171 Entries. One can open “Manuscript” facet to see all ms records associated with place.	In "Manuscripts" perspective: (A) Filter on "Production Place" by traversing the hierarchy to reach "Low Countries" - the result is 151. Then filter separately for Belgium (1,784), Flanders (2,413), Luxembourg (15), Netherlands (2,577), Southern Netherlands (307), Spanish Netherlands (1), Westhoek (19). Alternatively: (B) Use "Production Place" filter options and select "Bounding Box". Then draw a rectangle around the approximate area of the "Low Countries" - result will be something like 7,395.	(1) "Low Countries" is not a hierarchical region in the TGN vocabulary. You have to pick at least 8 different terms from the same level of the hierarchy under "Europe". (2) The Bounding Box can only be rectangular, and cannot follow the contours of geographical regions.

[H3] How many manuscripts were produced in London in the 15th century?	One can obtain a list of results by general browses on modern places (e.g. London) and then filter by the field “lieu” and by the object “livre”. But currently one cannot search by date range.	Can filter places of production by country and a few regions, but not by a specific city or town. Can add a century filter to countries of production.	Use Advanced Search to limit Manuscript Date to 1400-1501 and Production Place to London. 348 Entries. It is possible to get to only MS records associated with date, but it is not easy to count the faceted list of results.	In "Manuscripts" perspective: (A) use the "Production Place" filter to find "Greater London" in the hierarchy under "England". This gives 483 manuscripts. Then (B) use the "Production Date" time slider to limit the results to the 15th century. This gives 266 manuscripts.	(1) The "Production Place" timeslider is awkward to use with precision. The closest date range I could get was 1382 to 1495. (2) You might need to search the Production Places to find that "Greater London" is the correct term.
[H4] How many manuscripts formerly owned by Sir Thomas Phillipps are in British Libraries?	A general browse leads easily to Sir Thomas Phillipps and thus to his collection. In the record of his collection we find a list of manuscripts owned by him (currently just over 400, thanks to Synnøve’s work). This list is ordered by shelfmark and one can count by hand all British libraries.	Browse by “People” and select Thomas Phillipps. His role can then be filtered for “owner”. The result is a list of Oxford university and college MSS formerly owned by Phillipps.	This one is not really possible. The best you can do is limit search results to Entries with Phillipps as Provenance Agent. After that, you would have to go through every Entry/MS Record individually to determine which were last observed in British libraries.	In "Manuscripts" perspective: filter on "Owner" for Thomas Phillipps. This produces 8,752 manuscripts. Then filter on "Transfer of Custody Date" for dates after 1872 (death of Phillipps). This produces 916 manuscripts. You can then browse the list of owners (in the Owner filter panel on the left) for those which are British libraries. (Alternatively, you can filter additionally on	This is pretty clunky! It might be easier once "Last Known Location" has been implemented. (An alternative approach might involve starting with "Events", selecting "E7 Activity", and limiting to "date > 1872" + collection = "Phillipps". This produces 1,579 events which can be sorted and browsed by

				"Transfer of Custody Place" for places in Britain.)	manuscript label and/or date.)
[H5] What is the average number of folios in a book of hours?	One can do a general browse on “heures” or “d’heures”, and select the manuscripts in the result list. But there is currently no way of selecting or displaying their sizes without opening these records one by one.	Can find MSS which contain “book of hours”, but each language has a separate Work entry. Cannot then count or average the folio numbers but can browse folio counts of search results.	This is not feasible in the interface, but if you exported your search results to .csv you could arrive at an estimate. Limit results to those with Book of Hours (etc.) as Title, export results, then find the average folio count in .csv file.	In "Manuscripts" perspective: search under "Works" filter for "hours". This produces 4,319 manuscripts. The number of folios is displayed for each manuscript, where available. There is no way to find the average number of folios across the full list of manuscripts within the MMM interface itself. But if you export the results of the search and manipulate the resulting CSV file, you can find the average number of folios in that way. The answer is 159.86 folios (across a total of 3,899 manuscripts).	You would also need to screen out of the spreadsheet those manuscripts which are only fragments, not entire codices. You should also add "heures" and "horae" to the search.

[H6] Which collectors bought manuscripts from Wilfrid Voynich? Where were they located? What do we know about the kind of manuscripts he sold, and their earlier histories?	You can search for Voynich as a person, which links you to his collection and two of the MSS which were part of that collection. But you can't follow Voynich's activities as a seller of MSS.	Voynich does not appear in the Oxford catalogue, presumably because he neither sold nor owned any of the Oxford MSS.	Browsing the list of "Sellers" for Voynich produces 47 records. From there you can see the list of "Buyers" for 33 of these items. You can also see a list of the 52 works contained in these MSS. Browsing the "Provenance" facet for these items displays 17 previous or subsequent owners.	In the "Manuscripts" perspective, search for "Voynich" as an Owner. This produces a list of 213 MSS which can be browsed to see the other owners of each MS. To find more details of the other owners, you have to select each one separately.	The list of MSS includes Voynich acting in the roles of "collection owner" and "manuscript owner" as well as "selling agent".

Appendix 3: Publications by staff of the MMM Project

Antoine Brix (2019) "Reconstructing the Sorbonne Library in the Bibale Database: New Paths through Old Matter" https://libraria.hypotheses.org/1129
Eero Hyvönen, Esko Ikkala, Miho Koho, Jouni Touminen, Toby Burrows, Lynn Ransom and Hanno Wijsman, "A Linked Open Data Service and Portal for Pre-modern Manuscript Research", *Digital Humanities in the Nordic Countries 2019 Conference*, Copenhagen, March 2019 http://ceur-ws.org/Vol-2364/20_paper.pdf
Toby Burrows, Eero Hyvönen, Lynn Ransom, Hanno Wijsman, "Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts" *Manuscript Studies* 3 (1) (2018), 249-251
Toby Burrows "Connecting Medieval and Renaissance Manuscript Collections", *Open Library of Humanities*, 4 (2) (2018) 32 pp. DOI: http://doi.org/10.16995/olh.269
Toby Burrows "Tracing the History of Medieval and Renaissance Manuscripts: Two Recent Digital Humanities Projects", in: Lana Pitcher and Michael Pidd (eds), *Proceedings of the Digital Humanities Congress 2018*. Studies in the Digital Humanities. Sheffield: The Digital Humanities Institute, 2019. https://www.dhi.ac.uk/openbook/chapter/dhc2018-burrows
Burns, Halle; Burrows, Toby; Downie, J. Stephen; Lewis, David; Page, Kevin; Velios, Athanasios. 2019. "Assessing the practicality of ARK identifier usage in a catalogue of medieval manuscripts." *iConference 2019 Proceedings*. https://doi.org/10.21900/iconf.2019.103380
Toby Burrows, Antoine Brix, Douglas Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Synnøve Myking, Kevin Page, Lynn Ransom, Emma Cawlfield Thomson, Jouni Tuominen, Hanno Wijsman and Pip Wilcox. "Linked Open Data Vocabularies and Identifiers for Medieval Studies," *DHN 2020: Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, Riga, Latvia, October 21-23, 2020. Edited by Sanita Reinsone, Inguna Skadiņa, Anda Baklāne, Jānis Daugavietis. CEUR Workshop Proceedings, vol. 2612. pp. 211-218 http://ceur-ws.org/Vol-2612/short5.pdf
Burrows, T, Emery, D, Fraas, M, Hyvönen, E, Ikkala, E, Koho, M, Lewis, D, Morrison, A, Page, K, Ransom, L, Thomson, E, Tuominen, J, Velios, A and 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. DOI: https://doi.org/10.5334/johd.14

Toby Burrows, Athanasios Velios, Matthew Holford, David Lewis, Andrew Morrison and Kevin Page, "Transforming TEI Manuscript Descriptions into RDF Graphs", forthcoming 2020 - *GraphSDE proceedings*
Toby Burrows, Doug Emery, Arthur Mitchell Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Andrew Morrison, Kevin Page, Lynn Ransom, Emma Cawlfield Thomson, Jouni Tuominen, Athanasios Velios, Hanno Wijsman. "A New Model for Manuscript Provenance Research: the Mapping Manuscript Migrations Project". Submitted to *Manuscript Studies* April 2020

Appendix 4: Conference Presentations by staff of the MMM Project

Name of Event	Location	Date	Status	Disciplinary area
International Medieval Congress (Leeds) 2018	Leeds, UK	July 2018	Paper presented	Medieval and Renaissance studies
International Medieval Congress (Leeds) 2019	Leeds, UK	1-4 July 2019	Session presented	Medieval and Renaissance studies
International Medieval Congress (Leeds) 2020	Leeds, UK	July 2020	Paper and workshop originally accepted for conference; paper accepted for virtual conference	Medieval and Renaissance studies
International Congress on Medieval Studies (Kalamazoo, MI) 2018	Kalamazoo, US	May 2018	Paper presented	Medieval and Renaissance studies
International Congress on Medieval Studies (Kalamazoo, MI) 2019	Kalamazoo, US	May 2019	Workshop and paper presented	Medieval and Renaissance studies
Medieval Academy 2019	Philadelphia, US	7-9 March 2019	Paper presented	Medieval and Renaissance studies
Renaissance Society of America 2020	Philadelphia, US	April 2020	Proposal accepted - postponed	Medieval and Renaissance studies
CARMEN: Ghent	Ghent	September 2017	Paper presented	Medieval and Renaissance studies
ACMRS Conference 2018	Phoenix, AZ	February 2018	Paper presented	Medieval and Renaissance studies
Schoenberg symposium 2019 on linked data (digital and analog)	Philadelphia, US	November, 2019	Session presented (four presenters)	Manuscript studies
DH 2019 - Utrecht, June 2019	Utrecht	9-12 July 2019	Poster and workshop presented + paper on Oxford/Bodleian work	Digital Humanities Conferences

DHN 2018 (Digital Humanities in the Nordic Countries)	Helsinki	7-9 March 2018	Paper presented	Digital Humanities Conferences
DHN 2019 (Digital Humanities in the Nordic Countries)	Copenhagen	6-8 March 2019	Paper presented	Digital Humanities Conferences
DHN 2020 (Digital Humanities in the Nordic Countries)	Riga	20-23 October 2020	Proposal accepted - conference postponed	Digital Humanities Conferences
DH Benelux 2018	Amsterdam	June 2018	Accepted but had to withdraw	Digital Humanities Conferences
DCH - Digital Cultural Heritage	London	November 2017	Paper presented	Digital Humanities Conferences
“New Sources for Book History”, CERL/British Library, November 2017	London	November 2017	Paper presented	Library Conferences
ISMI: International Manuscript Identifier group - CERL/Liber/IRHT	Paris	October 2017	Attended	Manuscript studies
ISMI: International Manuscript Identifier group - CERL/Liber/IRHT	Paris	April 2018	Attended	Manuscript studies
Manuscript cataloguing meeting - November 2017 - Bodleian, Cambridge, British Library	Oxford	November 2017	Paper presented	Manuscript studies
Parker on the Web 2.0	Cambridge	March 2018	Paper presented	Medieval and Renaissance studies
Object Biographies	Helsinki	2-3 March 2018	Paper presented	Conferences
Linked Pasts 2018	Mainz	11-13 December 2018	Poster accepted	Digital Humanities Conferences
Digital Humanities Congress 2018	Sheffield	September 2018	Paper presented	Digital Humanities Conferences
Past, Present, and Future of Libraries	Philadelphia	27-29 September 2018	Paper presented	Library Conferences
Association for College and Research Libraries' Rare Books and Manuscripts Section	Baltimore	18-21 June 2019	Poster presented	Library Conferences
Workshop on Scholarly Digital Editions, Graph Data-Models and Semantic Web	Lausanne	3-4 June 2019	Paper presented	Digital Humanities Conferences

Technologies (GraphSDE)				
iConference 2019	Washington, DC	31 March - 3 April 2019	Poster presented	Library Conferences
Dark Archives (Medium Aevum journal / Society for Medieval Literature)	Oxford	Sept 2019	Paper presented	Medieval and Renaissance studies
Oxford/Cambridge Symposium on Manuscript Descriptions	Oxford	March 2019	Paper presented	Manuscript studies
DReAM Lab workshop on Linked Data	Philadelphia	10-14 June 2019	Paper presented	Semantic Web
HELDIG Summit 2017	Helsinki	October 2017	Paper presented	Digital Humanities
WHiSe: Workshop on Humanities in the Semantic Web III	Leipzig	22-23 May 2019	Paper submitted; workshop cancelled	Semantic Web
Digital Humanities (DIGIHUM) Academy of Finland Programme Annual seminar, November 2018	Helsinki	Novemb er 2018	Presented	Digital Humanities
Digital Humanities (DIGIHUM) Academy of Finland Programme Annual seminar, May 2017	Helsinki	May 2017	Presented	Digital Humanities
Ontologies workshop - DH 2019	Utrecht	8 July 2019	Presented	Digital Humanities Conferences
Materia on the Move workshop - DH 2019	Utrecht	8 July 2019	Presented	Digital Humanities Conferences
DIGIHUM Conference 2019	Tallinn	26 Sept 2019	Invited	Digital Humanities Conferences
Care and Conservation of Manuscripts 18	Copenhagen	22-24 April 2020	Proposal accepted - conference postponed	Manuscript studies
Digging into Data Challenge: end-of-grant conference	Alexandria, VA	29-31 January 2020	Paper presented	Digital Humanities
Digital Initiatives Symposium 2020	San Diego	28 April 2020	Concurrent session accepted - conference postponed	Digital Humanities
Linked Pasts 5 Conference : https://linkedpasts5.sciencesconf.org/	Bordeaux	11-13 Decembe	Poster presented	Digital Humanities

		r 2019		
"History of the Book" Seminar - Bodleian Library	Oxford	28/2/2020	Invited and presented	Manuscript studies
DH Benelux 2020	Leiden	3-5 June 2020	Proposal submitted - conference postponed	Digital Humanities
Data for History	Berlin	28-29 May 2020	Proposal submitted - conference postponed	Digital Humanities
Research Group on Manuscript Evidence	Princeton, NJ	13-14 March 2020	Invited - event cancelled	Manuscript studies
CILIP Cataloguing and Indexing Group Scotland	Glasgow	17 April 2020	Invited - symposium postponed	Library Conferences
London Rare Book School	London	18 June 2020	Invited - event postponed	Manuscript studies
EADH 2nd International Conference	Krasnoyarsk, Russia	22-25 September 2020	Paper submitted - conference postponed	Digital Humanities