Exploring a large graph of historical objects: the Mapping Manuscript Migrations project

Toby Burrows[1], Laura Cleaver[2], Doug Emery[3], Mikko Koho[4], Lynn Ransom[5], and Emma Thomson[6]

[1]University of Oxford; University of Western Australia
[2]University of London
[3,5,6]University of Pennsylvania
[4]University of Helsinki; Aalto University

1 Introduction

The Mapping Manuscript Migrations (MMM) project took data from three disparate sources relating to the history and provenance of Western medieval and renaissance manuscripts, and transformed them from TEI-XML documents or relational databases into a knowledge graph (Koho et al., 2021). The sources were the Schoenberg Database of Manuscripts (SDBM), a manuscript provenance database produced by the University of Pennyslvania Library; Bibale, a manuscript provenance database produced by the Institut de recherche et d'histoire des textes (IRHT); and Medieval Manuscripts in Oxford Libraries, produced by the Bodleian Library at the University of Oxford, which contains TEI-XML descriptions for more 10,000 Oxford manuscripts.

The resulting knowledge graph consists of more than 24 million RDF triples, together with a data model built on the CIDOC-CRM and FRBRoo ontologies, together with some MMM-specific entity classes and properties. The graph includes more than 222,000 manuscripts, 56,000 people and organizations, and 5,000 places, as well as 937,000 events and 435,000 works. These vocabularies, with the exception of events, were harmonized across the data sources, using a mixture of automatic and semi-automatic methods depending on whether the source datasets included reconciliable public identifiers from vocabularies like the Virtual International Authority File (VIAF), GeoNames, and the Getty Thesaurus of Geographical Names (TGN).

This paper examines and evaluates methods for traversing this knowledge graph, and for exploring, analysing, and visualizing the data. The MMM project implemented three major approaches to exploring the MMM knowledge graph: a public "semantic portal"; a downloadable and reusable copy of the entire MMM dataset; and a SPARQL endpoint. To test and evaluate different ways of traversing the MMM knowledge graph, the project developed a set of specific research questions, derived from consultation with focus groups, a comparable list from the Biblissima project, and the research interests of project members.

2 MMM Semantic Portal

The public MMM portal provides users with the capability to browse, search, filter, and download the data in the form of lists and CSV files. The Sampo-UI software used for the portal can also create map-based visualizations of the data to show the locations of manuscript production, together with last-known locations and the historical trajectories of manuscript ownership (Ikkala et al., 2021). The first iteration of the Sampo-UI portal was evaluated by using the initial set of research questions, leading to some modifications and some explanatory text and guidance for users of the portal, in the form of Frequently Asked Questions.

As the report on this evaluation made clear, some of the issues raised arose, not from the functionality of the portal per se, but from the different data models used by the three sources and from the ways in which the MMM data model was designed to accommodate these variations (Burrows et al., 2020). The "last-known location" property was preferred to "current location", for instance, primarily because the SDBM data consist of observations of specific manuscripts at specific dates in their history, but do not necessarily identify their current location, whereas the Oxford data describe manuscripts known to be in Oxford at the present day. Ambiguous and uncertain data in the sources also created difficulties with interpreting the result sets displayed in the MMM portal. The count of the number of manuscripts produced in Italy, for example, differed between the filtered listings by place and the map-based visualizations. This was the result of many manuscripts having more than one possible place of production.

3 Data Download and Reuse

The entire MMM data package, including the data model, can be downloaded from the Zenodo data repository as RDF Turtle files (Burrows et al., 2020). From there, the data can be imported into other software environments, which may enable different kinds of exploration and visualization. A total of 53 downloads had been recorded as of 20 November 2021.

The MMM project tested this process with the ResearchSpace software developed by the British Museum in partnership with metaphacts GmbH. We were able to import the entire dataset and successfully deployed some basic browsing and visualization functionality. The MMM data relating to more than 8,000 manuscripts formerly owned by the 19th-century British collector Thomas Phillipps were also exported into a nodegoat environment maintained by one of the project's investigators (Burrows), which enables browsing and filtering, as well as map and network visualizations.

4 SPARQL Queries

The third approach was to make a SPARQL endpoint available through the Linked Data Finland service. The MMM Project worked extensively with SPARQL through a series of weekly workshops over two years, which aimed to encourage Linked Open Data specialists to share their knowledge with the manuscript curators, researchers, and other technical staff involved. These sessions explored a series of increasingly complex research questions, which went beyond the initial set of queries used in modelling and testing the MMM data and evaluating the portal. An introductory MMM SPARQL tutorial, published on GitHub, was developed as a result of the workshops.

This process enabled a more diagnostic approach to be taken to the data, by using

```
SELECT
?production_year_average
(?height_mm_average / ?width_mm_average AS ?missal_ratio)
(?b_height_mm_average / ?b_width_mm_average AS ?breviary_ratio)

WHERE {
  {
    SELECT ?manuscript (AVG(?height_mm) AS ?height_mm_average) (AVG(?width_mm) AS ?width_mm_average) (AVG(?production_year) AS ?production_year_average)
    WHERE {
          ?manuscript a efrbroo:F4_Manifestation_Singleton ;
                      mmms:height ?height ;
                      mmms:width ?width ;
                      mmms:manuscript_work ?work .

     ?height ecrm:P90_has_value ?height_mm ;
             ecrm:P91_has_unit mmms:Millimetre .
     ?width ecrm:P90_has_value ?width_mm ;
            ecrm:P91_has_unit mmms:Millimetre .
     ?work skos:prefLabel ?work_label .
     FILTER (CONTAINS (LCASE (?work_label), "missal"))
     FILTER (?height_mm > 39 && ?height_mm < 500)
     FILTER (?width_mm > 39 && ?width_mm < 500)

     ?production ecrm:P108_has_produced ?manuscript ;
                 ecrm:P4_has_time-span ?production_time_span .
     ?production_time_span ecrm:P82a_begin_of_the_begin ?production_date .
     BIND (YEAR(?production_date) AS ?production_year)
     FILTER (?production_year >= 700)
  }
    GROUP BY ?manuscript
  } UNION {
      SELECT ?manuscript (AVG(?height_mm) AS ?b_height_mm_average) (AVG(?width_mm) AS ?b_width_mm_average) (AVG(?production_year) AS ?production_year_average)
    WHERE {
          ?manuscript a efrbroo:F4_Manifestation_Singleton ;
                      mmms:height ?height ;
                      mmms:width ?width ;
                      mmms:manuscript_work ?work .

     ?height ecrm:P90_has_value ?height_mm ;
             ecrm:P91_has_unit mmms:Millimetre .
     ?width ecrm:P90_has_value ?width_mm ;
            ecrm:P91_has_unit mmms:Millimetre .
     ?work skos:prefLabel ?work_label .
     FILTER (CONTAINS (LCASE (?work_label), "breviar"))
```

Figure 1: SPARQL query: height-to-width ratios for liturgical manuscripts.

SPARQL queries to understand the limits and gaps in the source datasets, as well as the effects of some decisions which had to be made in devising the MMM data model. Price data for manuscript sales, for example, only occur in the SDBM and are not sufficiently reliable or consistent to support analyses. Events in the history of individual manuscripts are not always described specifically enough in the source datasets to support detailed modelling and reasoning, with the result that more generic classes and properties have had to be used for many ownership relationships in the MMM data model. The difference between the current location of a manuscript and the more ambiguous concept of "last-known location" also affected the results of SPARQL queries.

The SPARQL sessions also enabled a more analytical and quantitative approach to the data, supporting such queries as comparing the height-to-width ratios of different types of liturgical manuscripts, or exploring varying ways of calculating and comparing the length of time for which manuscripts were retained in stock and re-advertised by different booksellers and dealers. Figure 1 shows an example of a SPARQL query designed to find the height-to-width ratios of two types of liturgical manuscripts: missals and breviaries. The results can be visualized with graphs, bubble charts, and scatterplots if a SPARQL query service like Yasgui is used (Rietveld and Hoekstra, 2017).

Figure 2 shows the results of this query as a scatterplot, representing a total of 12,169 manuscripts, with missals shown in blue and breviaries in red. Most of the manuscripts fall within the range 1:1 and 2:1, though the majority fall between 1.25:1 and 1.6:1. There is considerable similarity between the two different types. Relatively few manuscripts have ratios less than 1:1 (i.e., their width is greater than their height).

Figure 3 shows a SPARQL query designed to explore and compare stock retention by the rare book and manuscript dealers Bernard Quaritch and James Tregaskis. The "duration" refers to the number of days elapsing between the first listing of a manuscript in a sale catalogue and the last listing; each listing is defined as a "transfer

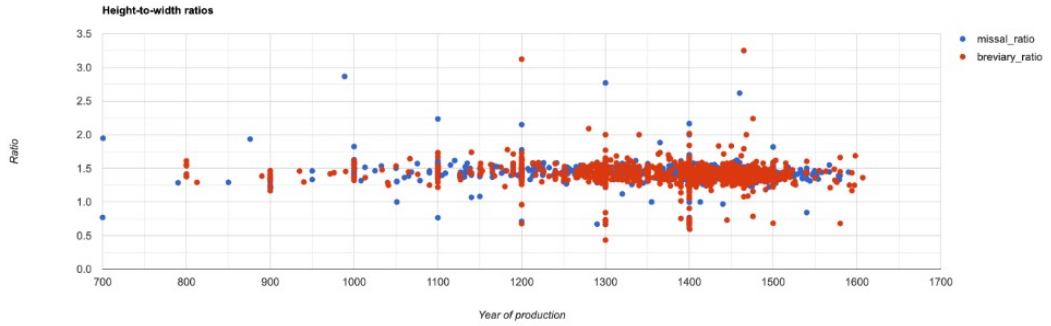

Figure 2: Scatterplot for height-to-width ratios of missals and breviaries.

```
SELECT ("" AS ?manuscript_name_sample) ?duration ?transfer_count ?seller (COUNT(?manuscript_name) AS ?manuscript_count)
  WHERE {
  {
SELECT ?manuscript_name (DAY(MAX(?timespan_datetime) - MIN(?timespan_datetime)) AS ?duration)  (COUNT(*) AS ?transfer_count) ?
seller
  WHERE {
  { ?transfer
    ecrm:P28_custody_surrendered_by | mmm:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_43130> ; #tregaskis
    ecrm:P30_transferred_custody_of | mmm:observed_manuscript ?manuscript .
     BIND ("tregaskis" AS ?seller)
  }
  UNION
  { ?transfer
    ecrm:P28_custody_surrendered_by | mmm:carried_out_by_as_selling_agent <http://ldf.fi/mmm/actor/bibale_30748> ; #quaritch
    ecrm:P30_transferred_custody_of | mmm:observed_manuscript ?manuscript .
    BIND ("quaritch" AS ?seller)
  }

    ?manuscript a efrbroo:F4_Manifestation_Singleton ;
    skos:prefLabel ?manuscript_name .

   ?transfer ecrm:P4_has_time-span ?transfer_timespan.
     ?transfer_timespan  skos:prefLabel ?timespan_label ;
                         ecrm:P82a_begin_of_the_begin ?timespan_datetime .

} GROUP BY ?seller ?manuscript_name
HAVING (?transfer_count > 1) (?duration > 0)
    )
} GROUP BY ?seller ?duration ?transfer_count
```

Figure 3: SPARQL query: Quaritch and Tregaskis comparison.

event".

In Figure 4, a selection of these events are visualized in a Yasgui bubble chart. Each bubble shows the duration, the number of transfer events, the seller (Tregaskis in red, Quaritch in blue), and the number of manuscripts with that combination of variables (in the size of the bubble). The most common combination is visible in the largest blue bubble in the lower left of the chart: a duration of 792 days and a transfer count of 2, with Quaritch as the seller. A total of 30 manuscripts have this combination. The configuration of the bubble chart has been used to limit the maximum duration shown to 5,000 days, for the sake of visibility.

The SPARQL queries also made it possible to enhance and extend the MMM data by combining endpoints for other data sources. Data about manuscript collectors from MMM can be enhanced from Wikidata, for example, with information about their occupations, places of birth, and gender. Figure ?? shows the results of a query of this kind. The simplest way of matching the person data from MMM with Wikidata was through VIAF identifiers, although this meant that persons without such identifiers were excluded from the results. A future desideratum would be an enhanced identity

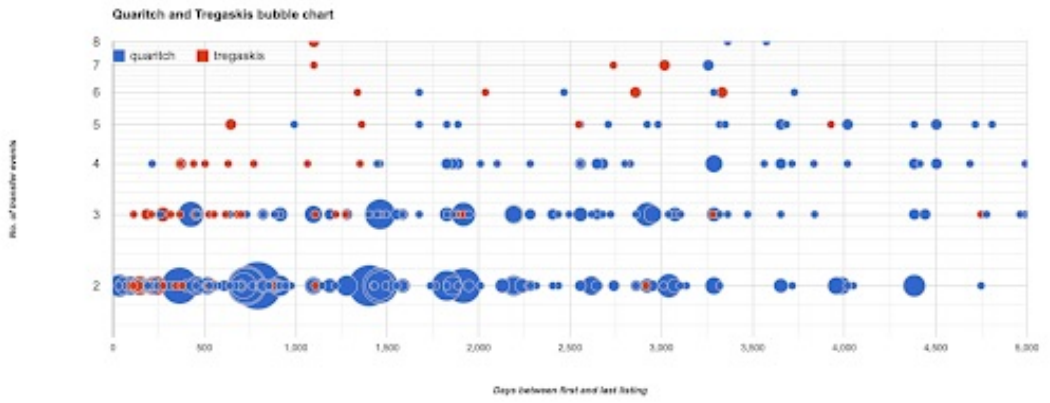

Figure 4: Bubble chart comparing Quaritch and Tregaskis.

resolution and reconciliation service for the main classes of entities, created either as a specifically medieval resource or by adding a wider range of identifiers to Wikidata.

SPARQL has had relatively little take-up by humanities researchers, even though a growing number of humanities services provide SPARQL endpoints to their data (Lincoln, 2015). But there are clear signs that being able to write SPARQL queries is becoming a useful practical skill for humanities researchers. The popular humanities data management, network analysis and visualisation environment nodegoat recently added functionality for using SPARQL queries to import contextual data from Linked Open Data sources, for example. SPARQL remains challenging to learn, even when using a detailed and well-documented data model like MMM, and requires a certain amount of trial and error.

The Yasgui interface used in the MMM workshop offers some diagnostic help with formulating queries correctly, but its main advantages are the built-in visualizations. Its new "Geo events" display which can produce timelines and map-based event sequences has also been tested against MMM data. But it would help to have a more visual approach to constructing the SPARQL queries themselves, in which data models and name spaces can be visualized for selecting entities and properties. One recent project has designed a visual interface for constructing SPARQL queries in the humanities, known as Gravsearch, but this has to be used within the Knora software package (Schweizer and Geer, 2021).

5 Conclusion

The three main approaches to exploring the MMM knowledge graph are intended to complement each other. The semantic portal provides users with sophisticated browsing, filtering, and searching across the aggregated data, together with some valuable visualizations. Being able to download the data provides other research projects with an important means of reusing the data in a different software environment or as part of a larger and more extensive service. Querying the SPARQL endpoint enables researchers and Linked Open Data experts to ask in-depth and complex questions, combine the data on the fly with other services, and perform diagnostic tests on the structure and content of the MMM knowledge graph.

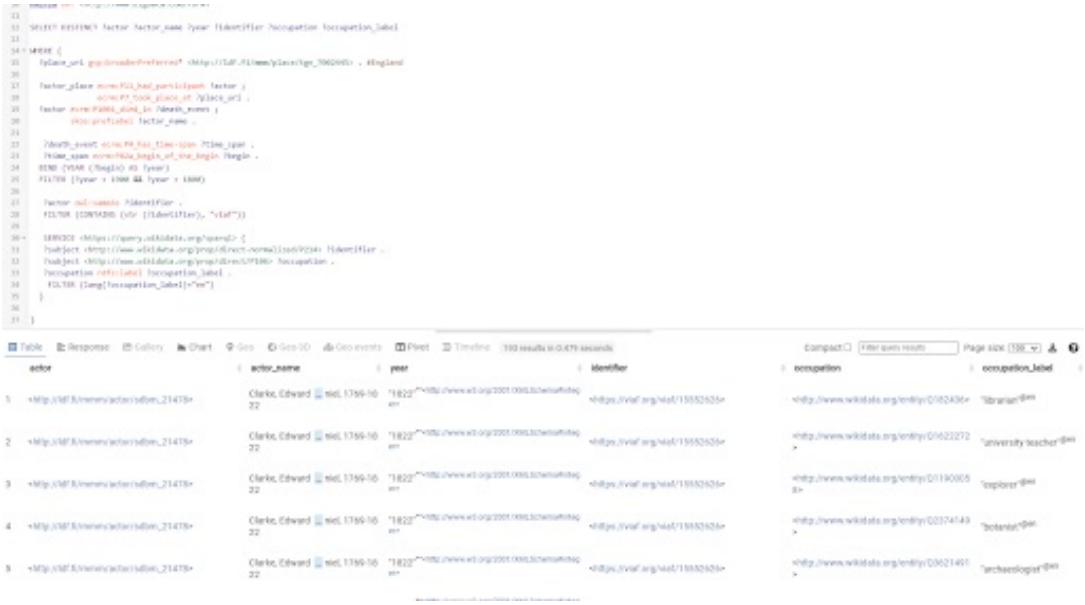

Figure 5: SPARQL query: persons enhanced from Wikidata.

Despite its considerable learning curve, the use of SPARQL queries can be a particularly valuable way of assessing the types of analysis and visualization which can be usefully applied to a knowledge graph, as well as for diagnostic exploration of the contents of the source data as reflected in the RDF transformations. One of the main lessons learned from the MMM project's experience with SPARQL has been the extent to which the scope and structure of the source datasets can impose limitations on quantitative and descriptive queries across the graph.

SPARQL can also demonstrate the ways in which gaps in the MMM data (in such areas as gender, occupation, and other biographical details) can be compensated for by drawing in this kind of information from other linked data sources. Being able to analyse and visualize data across multiple knowledge graphs, connected through URIs, is a major additional benefit of the Linked Open Data approach taken by projects like MMM.

References

Burrows, T., N. Bergk Pinto, M. Cazals, A. Gaudin, and H. Wijsman (2020). Evaluating a semantic portal for the 'Mapping Manuscript Migrations' Project. DigItalia: Rivista del Digitale nei Beni Culturali (2).

Burrows, T., D. Emery, M. Fraas, E. Hyvönen, E. Ikkala, M. Koho, D. Lewis, A. Morrison, K. Page, L. Ransom, E. Thomson, J. Tuominen, A. Velios, and H. Wijsman (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), 373–383.

Ikkala, E., E. Hyvönen, H. Rantala, and M. Koho (2021). Sampo-UI: a full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web 12(1), 1–16.

Koho, M., T. Burrows, E. Hyvönen, E. Ikkala, K. Page, L. Ransom, J. Tuominen, D. Emery, M. Fraas, B. Heller, D. Lewis, A. Morrison, G. Porte, E. Thomson, A. Velios, and H. Wijsman (2021). Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data. Journal of the Association for Information Science and Technology (JASIST).

Lincoln, M. (2015). Using SPARQL to access Linked Open Data. Blog. https://programminghistorian.org/en/lessons/retired/graph-databases-and-SPARQL.

Rietveld, L. and R. Hoekstra (2017). The YASGUI family of SPARQL clients. Semantic Web 8(3), 373–383.

Schweizer, T. and B. Geer (2021). Gravsearch: transforming SPARQL to query humanities data. Semantic Web 12(6), 379–400.